UNIVERSITY SOUTHERN QUEENSLAND

AN INVESTIGATION OF SKEWNESS, SAMPLE SIZE, AND TEST STANDARDISATION

A dissertation submitted by

Adina M. Piovesana B.Psych (Hons)

For the award of Doctor of Philosophy

2014

ABSTRACT

In psychological assessment, a raw score transformation is the first step in the clinical decision-making process. During this process, clinicians will transform a raw score typically using linear standardised scores and a normative sample with characteristics similar to their client. While the literature stresses the importance of using adequate normative data, little research has evaluated the effect skewness has on sample size. Currently the consensus is that a sample size of 50 is deemed adequate for normative data. However, an alarming number of studies that present normative data have much smaller sample sizes particularly when the data is stratified by age, gender, and/or education. Additionally, the use of linear transformations onto a normal distribution introduces further problems the more positively or negatively skewed the normative raw score distribution is. Skewed distributions are commonly encountered in neuropsychology and accordingly their deviation from a normal distribution should be considered during the clinical decision-making process. The primary goal of the current thesis was therefore to evaluate the psychometric issues related to the standardisation process. In particular to investigate the current understanding of sample sizes in neuropsychological samples, assess how this is influenced by different skewed distributions, and evaluate the potential errors involved in the decision-making process. Three studies were conducted. The first study explored the minimum sample size needed to produce stable measures of central tendency and variance for a range of distributions. Results indicated that the optimal sample size required was dependent on the level of skewness of the distribution and was not the often cited N = 50. For normally distributed data, a sample size of 70 is required in each cell in order to produce stable means and standard deviations. Negatively or positively skewed distributions required sample sizes that ranged from 30 to 80 in each cell. This study highlighted the inadequacy of currently available normative data and called for further normative research to be conducted. The second study evaluated the errors introduced when using three different linear transformations on different skewed distributions with adequate sample sizes. Seven tests with differing skewness coefficients were evaluated using the *z* score transformation, a *t*-test method developed by Crawford and Howell (1998), and a median *z* score transformation developed for this research. Results indicated that the traditional *z* score transformation produced the least errors of the three methods. However, for highly positively skewed distributions, the use of this transformation introduces considerable error in the clinical decision-making process. A regression equation was derived as a tool for clinicians to help correct adequate data for the effect of skewness. The final study evaluated whether using different linear transformations created substantial errors when using normative data that ranged in skewness and that had sample sizes less than those recommended from Study One. This study is particularly important given the common practice in neuropsychology for at least some measures to be derived from the clinical research literature utilising inadequate sample sizes. Results indicated that the error in judgement when using the preferred *z* score transformation is nearly doubled in positively skewed distributions. It was recommended that normative data with sample sizes less than 30 should not be used in clinical practice and guidelines were proposed for incorporating issues of sample size and skewness into their testing practices. It is hoped that clinicians will adopt the findings and subsequent recommendations of these studies in order to improve the current standards of clinical decision-making in neuropsychology.

CERTIFICATE OF DISSERTATION

I certify that the ideas, experimental work, results, analyses and conclusions reported in this dissertation are currently my own effort, except where otherwise acknowledged. I also certify that the work is original and has not been previously submitted for any other award except where otherwise acknowledged:

_____ _____

Signature of Candidate Date

ENDORSMENT

_____ _____

Signature of Supervisor Date

ACKNOWLEDGEMENTS

My gratitude firstly has to go to Graeme. You have both supported and mentored me through my undergraduate degree, my registration as a psychologist, and through my never-ending Doctoral degree. You have been my inspiration and the reason why I am so passionate about neuropsychology and all things brain. Thank you for all your time and effort, and the marvellous conversations.

Hannie. You and I started on this journey nine years ago and you have been my rock, my light and everything in between. I would never have accomplished this feat without your support, the endless pots of herbal tea, and the study sessions on your floor with Bells. Thankyou to your loving parents, Gerry and Mary for the hospitality and for treating me like their own.

My parents and family. Thank you for allowing me to live at home rent-free during my PhD. Words can never explain how thankful I am of your support and love.

Lastly to Darren, you joined me on this journey towards the end but it was at the most critical stage. You pushed me when I wanted to give in and never stopped believing in me. Your endless love, praise and support will stay with me forever. I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FORMULAE

CHAPTER ONE
INTRODUCTION

1.1 Introduction

The discipline of neuropsychology is concerned with the relationship between human brain function and behaviour (Beaumont, 2008). Neuropsychological assessment is the process of evaluating cognitive and psychosocial functioning in relation to neuropathology and how it influences an individual's ability to function in everyday life (Goldstein, 2005). The initial level of clinical evaluation involves comparing the individual's score from a test with a set of normative data that is representative of the *normal population*. These inferences are only made after the raw score is transformed into a standardised score and are compared with this normal population, a process called *standardisation.* Whilst this may be common knowledge for clinicians, Donnell, Belanger, and Vanderploeg (2011) note:

> "In the clinical practice of neuropsychology it is imperative to understand the psychometric properties of the measures used to evaluate patients and how those properties might affect decision-making about individual patients. Neuropsychological test performance has little meaning without understanding how an individual's score compares to the normative sample." (p. 1097)

These authors make a point that appears to be often disregarded or ignored in neuropsychology, the issue of standardisation. Much of the psychometric and clinical research literature has focused primarily on the issues of reliability and validity and how these influence neuropsychological assessments. Through clinical training and numerous neuropsychology textbooks, clinicians are taught the importance of considering the appropriateness, adequacy, reliability, and validity of any psychological test being administered to a client. In addition, professional ethics codes in psychology require clinicians to use psychometrically reliable and valid measures in clinical practice (Australian Psychological Association, 2007; Canadian Psychological Association, 2000; American Psychological Association, 2010). It is surprising, therefore, that only basic standardisation information is taught and provided to clinicians. As an example, in the *Handbook of Psychological Assessments – Fourth Edition* (Groth-Marnat, 2003) some discussion is based around general standardisation issues such as adequacy of the sample size, standardised administration, and the appropriateness of the normative data. While these issues are fundamental in test interpretation, the authors do not provide any guidance on how to address these issues. Rather, they are questions posed to the reader requiring them to consider these issues when providing an assessment.

In the *Handbook of Normative Data for Neuropsychological Assessment* (Mitrushina, Boone, Razani, & D'Elia, 2005) these standardisation issues are discussed in more depth. The authors provide a chapter on statistical and psychometric issues covering the standardisation of raw scores, the normal distribution, reliability, validity, and meta-analysis. While they introduce some issues, they do not provide evidence-based practice guidelines or consider the consequences of not contemplating the psychometric issues related to standardisation. Scrutiny of these issues is the purpose of the current thesis, which was conducted with the aim of providing recommendations, solutions, and theoretical guidelines relevant to the practice of everyday clinical neuropsychology.

It is hoped that better awareness of these issues will transform into more evidence-based and informed clinical judgements and subsequently improvement in the assessment and care of clients.

1.2 Neuropsychological Assessments

The psychometric issues of standardisation are not only applicable to individual test scores but also to scores within a test battery which consists of multiple measures that provide data over a broad range of cognitive and psychosocial domains (Vanderploeg, 1994). There are two predominant approaches to neuropsychological test batteries, with most clinicians fitting somewhere on the continuum between "fixed" and "flexible" testing (Vanderploeg, 1994). Fixed batteries use an unvarying group of tests that are administered in their entirety and are designed to comprehensively cover a broad range of cognitive domains in order to identify possible deficits (Lezak, 1995). They are also customarily co-normed and standardised on a single sample. The Halstead-Reitan Neuropsychological Battery (HRNB; Reitan & Wolfson, 1985) is perhaps the most well known example of a fixed battery. On the other end of the continuum is the flexible approach in which clinicians choose and interchange tests for each individual with the resulting battery tailored to the needs of each client (Vanderploeg, 2000). This clinically oriented approach allows clinicians to select tests based on hypotheses generated through clinical interviews, referral questions, and behavioural observations (Cimino, 2000). In addition to these more formal reasons, tests may also be selected based upon availability, routine, or clinical training.

While there are many advantages to the fixed battery, this approach also suffers from a number of major limitations. Because a fixed battery is administered in its entirety regardless of the status of the client, tests that measure a cognitive domain of no relevance to the case may be administered. Similarly, if one of the tests indicates that a particular cognitive domain is intact, the remaining related measures must still be administered. This means that such assessments can be extremely time-consuming, inefficient, and expensive.

Despite the appeal of a comprehensive fixed battery, research indicates that most clinicians adopt an intermediate approach, in which a core group of tests are repeatedly administered with other tests being added or substituted as needed (Sweet, Nelson, & Moberg, 2006; Sweet, Moberg, & Suchy, 2000; Sweet, Moberg, & Westergaard, 1996; Sweet & Moberg, 1990). This approach is clearly flexible. By using some aspects of fixed batteries and including smaller tests that supplement relevant cognitive domains, clinicians are able to integrate both approaches. What most clinicians fail to recognise is that such a "semiflexible" approach may undermine the psychometric properties of the fixed battery approaches which they seek to emulate. Without proper psychometric evaluation of the measurement error, reliability and validity associated with the individual tests and the battery as a whole, clinicians may be using combinations of tests that are psychometrically unstable, and may make test-based inferences of an individual's cognitive functioning based on error (Ingraham & Aikken, 1996).

The doctoral research of Olm-Madden (2008) conducted at the University of Southern Queensland compiled the mathematical procedures for computing the necessary psychometric properties in a flexible battery in an approach termed the Reliable Approach to Psychological Testing (RAPT). This method allowed "the application of psychometric, actuarial methodology to a flexible collection of cognitive tests" (Olm-Madden, 2008, pg. 3) and evaluated the psychometric issues

related to them. The RAPT methodology permits an extensive evaluation of the implications and consequences of test selection and substitution on the reliability of batteries and test combinations, essentially permitting the type of analysis found in fixed batteries to be applied to flexible batteries. While this is an invaluable tool for clinicians and researchers alike, RAPT does not consider standardisation issues and essentially focussed on the incorporation and use of psychometric characteristics without reference to the standardisation samples from which they were derived. Consideration of the issues associated with using standardised scores and normative data is the focus of the current research.

1.3 Summary of Thesis

The structure of this thesis will be to review relevant psychometric literature and highlight critical issues at each step of the standardisation process. Each issue will be evaluated through studies, and recommendations and methodologies for use in clinical practice will be provided.

Chapter Two discusses the basic standardisation concepts that are employed by clinicians in clinical neuropsychology. This includes discussing the different ways of standardising a raw score, the ongoing debate surrounding the use of percentile ranks in neuropsychology, and an introduction to normal and skewed distributions.

Chapter Three introduces normative samples, samples that are theoretically representative of the population. Also evaluated are the different methods used by clinicians and researchers to collect normative samples, and the representativeness of these samples for the individual or client being tested. Finally, this chapter addresses and analyses the issue of sample size and its relation to establishing stable means and standard deviations for different distributions (i.e., normal and skewed). Recommendations are provided to aid clinicians and researchers.

Chapter Four focuses on the different approaches commonly employed by clinicians when deciding whether a standardised score is abnormal. This decision-making process is important on two levels. The first level is concerned with the individual test score and is normally interpreted with the aid of abnormality cut-off scores. The second level is assessing abnormality of the difference between two test scores.

Chapter Five is concerned with evaluating the effects of skewness on standardisation. In particular, this chapter empirically evaluates the errors produced when three linear transformations are applied to a range of skewed distributions for tests commonly used in neuropsychology. This chapter also integrates the information from Chapter Three and assesses the effect of skewness and sample size on the clinical decision-making process. Recommendations for clinical practice are provided.

Chapter Six summarises the main findings of the thesis. These findings are evaluated in terms of common practice, and recommendations are made to aid clinicians in the clinical decision-making process. The implications to the discipline of neuropsychology are also explored.

CHAPTER TWO
BASIC PSYCHOMETRIC CONCEPTS

2.1 Introduction

In order to fully understand the issues of standardisation and normalisation most relevant to clinical interpretation, it is necessary to understand the related and basic psychometric properties. Whilst most of these concepts are common knowledge in psychology, there is considerable scientific debate over the proper use of standardised scores and the existence of percentiles in cognitive reports. That said, the way clinicians utilise and process normative data introduces and/or masks errors that have an effect on clinical decision-making

2.2 Basic Psychometric Concepts

When cognitive tests are administered, the operational value that is obtained after scoring is called the *raw score*. Depending on the test itself, the raw score can be represented as a rating, a function of time, or the number of correct or incorrect items (Gregory, 2007). In this form, the raw score tells the clinician little about the individual's ability. Inferences about the individual's performance are only achieved after the raw score is compared with others of similar characteristics on the same cognitive test (Gregory, 2007). This comparison group is referred to as the *normative sample*, and is intended to be representative of the *population*. However, in order to complete this comparison, the clinician must convert the raw score into the meaningful *percentile rank* and/or *standardised score*.

2.2.1 Percentile rank

In its simplest form, percentile ranks are the ordinal positions of raw scores within a normative sample's distribution (Mitrushina et al., 2005). For example, a raw score of 123 on the Peabody Picture Vocabulary Test, Fourth Edition for an eight-year-old boy, equates to the 37^{th} percentile. That is, this score is as good as or better than 37 percent of people in the normative sample. The advantage, as described by Crawford and Garthwaite (2009) is that "percentile ranks express test scores in a form that is of greater relevance to the neuropsychologist than *any* alternative metric because they tell us directly how common or uncommon such scores are in the normative population" (p. 194). Some also go further to highlight that percentile ranks are universally applicable (Anastasi & Urbina, 1997) and can be readily understood not only by clinicians but also by other professionals and clients themselves (Lezak, Howieson, Loring, Hannay, & Fischer, 2004).

However, as a percentile rank is the relative standing of a score in a normal distribution, it does not provide any information on the difference between scores (Mitrushina et al., 2005). Additionally, differences between the percentile ranks on the distribution are wider towards the mean or the median and narrower at the upper and lower limits of the distribution (Anastasi & Urbina, 1997). As such, it is difficult for the clinician to determine whether a difference between the 5^{th} and 15^{th} percentile is bigger than a difference between the 55^{th} and 65^{th} percentile. Another major disadvantage of percentile ranks is that they have little use in the combination process for decision-making. That is, it is difficult to combine different percentile ranks for a variety of cognitive tests and determine an individual's strengths and weaknesses.

Irrespective of these major limitations, percentile ranks also present many advantages, especially for clinical neuropsychology. One of the assumptions so far

has been that the distribution of the normative sample, from which percentile ranks are derived, is based around *normality* or a distribution that is *symmetrical*. It is important to note that this is not always the case. Because cumulative percentiles are based on the actual raw score distribution, if a distribution is *skewed* or *asymmetrical,* as is the case with many neuropsychological tests, then interpretative power will not be lost (Donnell, Belanger, & Vanderploeg, 2011; Brooks, Strauss, Sherman, Iverson, & Slick, 2009). In other words, the interpretive power of these percentile ranks is not affected by skewness in the normative sample (Crawford, Garthwaite, & Slick, 2009).

Another issue that warrants investigation is the three different ways to conceptualise a percentile rank. Crawford et al. (2009) discuss this issue in depth and describe three definitions as follows:

1. The percentage of scores that fall below a given score
$$(m/N) \qquad \textit{Formula 1}$$
2. The percentage of scores that fall at or below a given score
$$(m + k)/N \qquad \textit{Formula 2}$$
3. The percentage of scores that fall below a score and half of those obtaining the score of interest
$$(m + 0.5k)/N \qquad \textit{Formula 3}$$

Where:

m = the number of people scoring below the given score
k = the number of people obtaining the given score
N = the overall size of the normative sample

While these authors appreciated that the three percentile definitions may cause minimal differences, they do highlight that such differences may be greater when tests consist of a small number of items, have a normative sample with a small sample size, or when the normative sample distribution is skewed. They emphasise this point with a worked example. The authors firstly created a hypothetical frequency distribution of raw scores on a 12-item neuropsychological test, and then calculated the percentile ranks of the raw scores using each of the three methods for a normative sample of 100 people. Table 1 reproduces the results of the Crawford et al. (2009) study.

Table 2.1.
Applying Three Different Definitions of a Percentile Rank to the Raw Scores

Raw Score	*n* obtaining	Percentile Ranks		
		Definition A: m/N	Definition B: $(m + k)/N$	Definition C: $(m + 0.5k)/N$
0	0	<1	<1	<1
1	0	<1	<1	<1
2	0	<1	<1	<1
3	0	<1	<1	<1
4	0	<1	<1	<1
5	2	<1	2	1
6	4	2	6	4
7	4	6	10	8
8	14	10	24	17
9	16	24	40	32
10	20	40	60	50

11	30	60	90	75
12	10	90	>99	95

As can be appreciated, the effects of applying the three different methods are considerable. For example, for a raw score of 9 (N = 16), definition A would provide a percentile rank of 24; definition B would provide a percentile rank of 40; and definition C would provide a percentile rank of 32. Using the qualitative classifications presented in Table 2, definition A would yield a description in the Low Average range whilst definitions B and C would result in an Average range classification.

Table 2.2.
Qualitative Classifications used in Neuropsychology

Classification	Lower Limit Percentile
Extremely High*	98
Above Average	91
High Average	75
Average	25
Low Average	9
Below Average	2
Extremely Low*	-

** The terms Extremely High and Low have replaced Significantly Above and Below Average*
Adapted from Kramer (1990)

Although this debate may divide the scientific community, it begs the question to what extent does this affect the clinical decision making process? Although Crawford et al. (2009) have clearly demonstrated the differences between the three methods, are these differences significant enough to influence the overall process? Study One critically evaluates this issue below.

2.2.1.1 Study One - Different Percentile Definitions

The different percentile ranks used in the Crawford et al. (2009) study have been utilised, but using actual normative data rather than a hypothetical data set and from four tests with 30 (Judgement of Line Orientation; JLO), 38 (Conceptual Level Analogy Test; CLAT), 50 (National Adult Reading Test; NART), and 60 (Boston Naming Test; BNT) items in an effort to examine the extent of the differences with larger item sets. A test with only 12 items would, by necessity, generate a discontinuous set of percentile ranks and accordingly magnify the apparent differences between the percentile methods.

Participants

The normative data for each of these tests was drawn from ongoing normative studies conducted through the Department of Psychology at the University of Southern Queensland. Ethics approval was obtained through the University of Southern Queensland (H10REA096). Participants were generally from regional South-East Queensland or metropolitan Brisbane areas and had volunteered to participate in studies designed to establish Australian norms for a number of neuropsychological tests. Table 2.3 presents the descriptive statistics for each of the normative studies used in this worked example.

Table 2.3
Demographic Characteristics for the Four Normative Samples

				Age	Education	Gender	
Test	*n*	*M*	*SD*	*M (SD)*	*M (SD)*	M	F
BNT	571	54.5	3.83	37.18 (14.60)	12.67 (2.39)	245	326
CLAT	265	24.00	6.91	33.13 (13.63)	12.46 (2.10)	101	164
JLO	379	25.61	3.67	35.90 (16.05)	12.50 (2.39)	150	229
NART	160	30.47	8.54	37.76 (15.47)	13.31 (2.27)	68	92

Materials

Four tests were utilised in this study. The Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 1983) is a confrontational naming test that requires the participant to name pictures, ordered with increasing difficulty. It is used to assess word retrieval abilities and word-finding difficulties. The BNT was originally developed in 1978 for use in adult populations but has since been normed for children and was revised in 1983 to include 60 items instead of the original 85. The Conceptual Level Analogy Test (CLAT) is a 42-item multiple-choice analogy test that assesses abstract reasoning. It was adapted by Willner in 1971 from the original Willner-Scheerer Analogy Test (1965). The Judgement of Line Orientation (JLO; Benton, Varney, & Hamsher, 1978) is a test designed to assess spatial perception and orientation. It consists of 30 items of increasing item difficulty. The participant must identify which line is in the exactly same position and orientation as a target item. Lastly, the National Adult Reading Test (NART; Nelson & O'Connell, 1978) is a 50 item reading test used as a measure of premorbid functioning.

Tables 2.4 to 2.7 display the percentile ranks of the raw scores using each of the three methods for the normative samples.

Table 2.4.
Percentile Ranks for Three Different Definitions for the Judgement of Line Orientation Test

Items	1-4	5-13	14	15	16	17	18	19	20	21	22	23	24	25	26
n obtaining	0	1	1	3	4	3	7	11	14	12	8	23	25	40	41
Definition A	0	<1	<1	<1	1	2	3	5	8	12	15	17	23	30	40
Definition B	0	<1	<1	1	2	3	5	8	12	15	17	23	30	40	51
Definition C	0	<1	<1	<1	2	3	4	6	10	13	16	20	26	35	46

Items	27	28	29	30
n obtaining	41	57	55	33
Definition A	51	62	77	91
Definition B	62	77	91	100
Definition C	56	69	84	96

Table 2.5.
Percentile Ranks for Three Different Definitions for the Conceptual Level Analogy Test

Items	0-2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
n obtaining	0	1	1	0	1	3	2	1	3	2	4	0	4	4	8
Definition A	0	0	<1	<1	<1	1	2	3	4	5	5	7	7	8	10
Definition B	0	<1	<1	<1	1	2	3	3	5	5	7	7	8	10	13
Definition C	0	<1	<1	<1	<1	2	2	3	4	5	6	7	8	9	11

Items	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
n obtaining	9	7	13	14	10	14	17	21	14	14	11	17	10	12	14
Definition A	13	16	19	24	29	33	38	44	52	58	63	67	73	77	82
Definition B	16	19	24	29	33	38	44	52	58	63	67	73	77	82	87
Definition C	14	18	21	26	31	35	41	48	55	60	65	70	75	79	84

Items	32	33	34	35	36	37	38	39	40	41	42
n obtaining	10	6	4	5	1	3	2	2	1	0	0
Definition A	87	91	93	94	96	97	98	98	99	100	100
Definition B	91	93	94	96	97	98	98	99	100	100	100
Definition C	89	92	94	95	96	97	98	99	99	100	100

Table 2.6.
Percentile Ranks for Three Different Definitions for the National Adult Reading Test

Items	1	2-4	5	6-12	13	14	15	16	17	18	19	20	21	22
n obtaining	0	1	1	0	3	2	0	3	1	4	3	2	2	2
Definition A	0	<1	<1	1	1	3	4	4	6	7	9	11	13	14
Definition B	0	<1	1	1	3	4	4	6	7	9	11	13	14	15
Definition C	0	<1	<1	1	2	4	4	5	7	8	10	12	13	14

Items	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
n obtaining	6	5	4	6	9	12	7	7	10	4	7	7	5	6	5
Definition A	15	19	22	24	28	34	41	46	50	56	59	63	68	71	75
Definition B	19	22	24	28	34	41	46	50	56	59	63	68	71	74	78
Definition C	17	20	23	26	31	38	43	48	53	58	61	65	69	73	76

Items	38	39	40	41	42	43	44	45	46	47	48	49	50
n obtaining	7	8	1	5	2	3	3	3	2	1	0	0	1
Definition A	78	82	87	88	91	92	94	96	98	99	99	99	99
Definition B	82	87	88	91	92	94	96	98	99	99	99	99	100
Definition C	80	84	87	89	91	93	95	97	98	99	99	99	100

Table 2.7.
Percentile Ranks for Three Different Definitions for the Boston Naming Test

Items	1-36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
n obtaining	0	1	0	0	0	3	2	2	3	7	6	11	11	15	18
Definition A	0	0	<1	<1	<1	<1	<1	1	1	2	3	4	6	8	11
Definition B	0	<1	<1	<1	<1	<1	1	1	2	3	4	6	8	11	14
Definition C	0	<1	<1	<1	<1	<1	<1	1	2	3	4	5	7	9	12

Items	51	52	53	54	55	56	57	58	59	60
n obtaining	32	34	44	45	71	69	79	53	35	30
Definition A	14	19	25	33	41	53	66	79	89	95
Definition B	19	25	33	41	53	66	79	89	95	100
Definition C	17	22	29	37	47	59	72	84	92	97

As can be seen from this example, the differences between the three methods are minimal in the grand scheme of clinical decision-making. The mean difference between Definition A and Definition B, the most discrepant methods, is 3.30 percentile points for the JLO, 2.34 for the CLAT, 1.98 for the NART, and 1.65 for the BNT. The differences are also minimal within the lower limits of the distribution. This is particularly important, as clinicians are primarily interested in the extreme ends of the distribution as more indicative of abnormality.

Overall, however, many researchers have concluded percentile ranks still hold great value in neuropsychological assessment and should be used alongside the standard metric of a standardised score (Crawford et al., 2009; Crawford & Garthwaite, 2009). As such it is important to report which method is being used when referring to percentiles. In the context of the current study, it is definition B that is utilised in which the percentile rank is computed as the percentage of people who score at or below the indicated score.

2.2.2 Standardised Scores

Raw scores do not contain any interpretive power without being converted into a percentile rank or a standardised score. Additionally, it is difficult to combine raw scores from different tests within a cognitive battery because they each present with different weightings and actual distributions. Standardising a score resolves these problems by transforming each raw score onto a common scale, which allows measures to be combined and analysed using known mathematics (Mitrushina et al., 2005; Lezak, Howieson, & Loring, 2004). This was a major limitation of percentile ranks and one that has been addressed in "fixed" batteries. The co-standardised methodology of scaled scores (SS) found in the "fixed" Wechsler scales (Wechsler, 1981) in theory allows an accurate comparison between standardised scores. However, for many tests used in the "flexible" battery approaches, clinicians need to complete the standardisation process themselves. Standardised scores can be represented in two distinct ways, *linear scores and normalised scores*, each with their own advantages, limitations, and psychometric issues. In order to discuss each type of standardised score, it is firstly important to understand *normal* and *skewed distributions* and the concept of the *normative sample*.

2.2.3 The Normal Distribution

The normal curve, bell curve, or Gaussian distribution was a theory first published by Carl Friedrich Gauss in 1809 in his book titled "Theory of the Motion of Celestial Bodies Moving Around the Sun in Conic Sections" (Davis, 1857). The general formula for the normal distribution curve is presented below in Formula 4 (Guilford, 1936):

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \qquad \textit{Formula 4}$$

Where N = the number of measurements

e = the base of the Naperian system of logarithms, 2.718

π = pi, or 3.1416

σ = sigma, the standard deviation of the distribution

x = a deviation from the mean ($X - M$)

The theoretical mean of the normal distribution is zero and the standard deviation is one. As Guilford (1936) explains:

> "The first terms, *N*, σ, and the square root of 2π, are constant for any distribution. They have nothing important to do with the general shape of the curve. The symbol *e* is also a constant value, namely, 2.718. The independent variable *x* appears in the exponent of the number 2.718. *Y* changes according to that exponent, and the value of the exponent changes according to the value of *x*. Let us assign a few values to *x* and then see what happens to *Y***.** If *x* is equal to zero, the whole exponent becomes zero. We know that any number to the power zero is equal to 1, no matter what the number may be. Thus *e* to the power of 0 equals 1. We know from this fact that the expression $e_{2\sigma^2}^{-x^2}$ will never be greater than 1 and that, when *x* departs from zero, either plus or minus, this expression becomes smaller. The curve will be symmetrical around the Y-axis because of the x^2 in the equation (p. 85)."

Figure 2.1 depicts the normal curve distribution with its characteristic "bell-shaped" curve and symmetry.

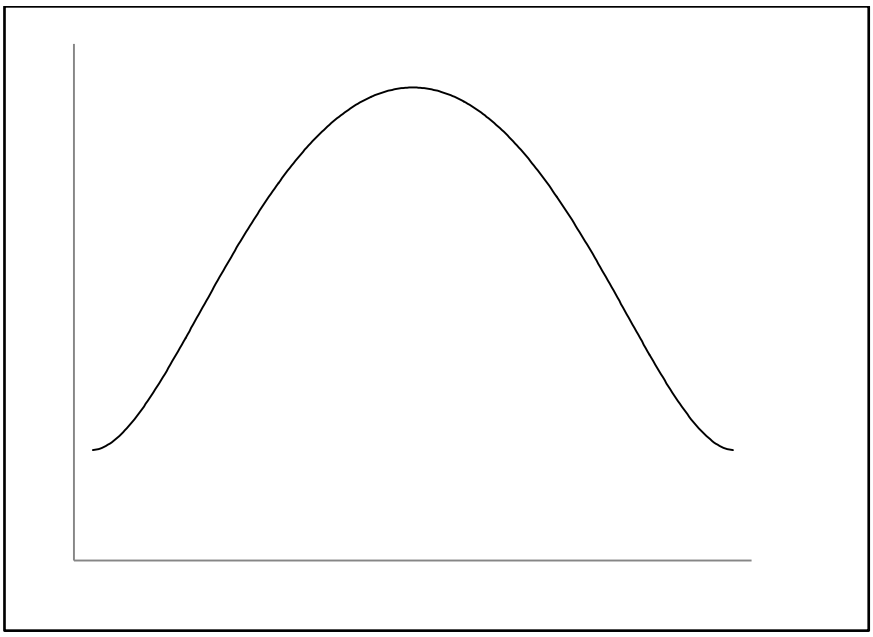

Figure 2.1. The normal distribution

The characteristic of the normal distribution that is commonly utilised by clinicians is that the area under the curve equals one. The normal curve is symmetrical and has the highest frequency of scores falling around the middle of the distribution. As such, the mode, median, and mean of the distribution are identical. This means that 68.4 percent of the population falls within one standard deviation of the mean. Essentially, this allows clinicians to make inferences about how a score compares to the population. An advantage of the theoretical normal distribution over observed distributions is that the former is mathematically defined and consequently we know essentially everything about it.

Overall, the normal distribution simplifies the representation of the world around us. Few human behaviours are likely to be truly normally distributed, and without transformation to a mathematically known distribution, researchers would have to know about the actual distribution of behaviours before making any inferences regarding them. For example where time is a dependent variable, there will be physical limitations on how quickly an individual can respond but no theoretical limitation on how slow. Transforming the data onto a normal distribution

allows us to make theoretical inferences about the meaning of particular scores despite not knowing the exact parameters of the positively skewed distribution. Most human characteristics and behaviours can be expressed using the normal curve (Mitrushina et al., 2005). The clinician can then take any portion of the normal curve and know what the corresponding percentage is for the area under the curve. This underlies the transformation of *linear standardised scores*.

2.2.4 Linear Standardised Scores

The *z* score is the most commonly used method by clinicians for converting raw scores into standardised scores (Bridges & Holler, 2007). The *z* score is evaluated by consulting the area under the normal curve and can be expressed by the following formula:

$$z = [x - M]/ SD \qquad \text{Formula 5}$$

Where:

x = the observed score
M = the mean
SD = the standard deviation

The benefit of a linear transformation is that the differences between the standardised scores are comparable to the differences in the equivalent raw scores. By and large, this allows the clinician to calculate the differences between scores, a major limitation of the percentile rank (Mitrushina et al., 2005). The main problem of using *z* scores is that they are expressed as both negative and positive numbers with decimal places, with negative scores falling below the mean, and positive scores above the mean. It is likely that most clinicians do not report *z* scores in their reports for this very reason and also because negative numbers can be difficult to compute. To resolve this issue, a number of scaling systems have been developed to represent these same values. These scaling systems are identical to *z* scores, but have adjusted means and standard deviations so that they are expressed as positive whole numbers (Anastasi & Urbina, 1997). The underlying distribution is not affected by this process. For example, T-scores have a mean of 50 and a standard deviation of 10. The formula for a T-score is:

$$T = 10z + 50 \qquad \text{Formula 6}$$

Where:

z = the z score

Other commonly used scaling systems are the scaled scores (ss) with a population mean of 10 and a standard deviation of three, and the Standard Score (SS) with a mean of 100 and a standard deviation of 15. Any linear transformation can be computed to any scaling system with the following general formula:

$$\text{Scale Score} = \text{Scale standard deviation } (z) + \text{Scale mean} \qquad \text{Formula 7}$$

The benefit of assuming (or asserting) normality of behaviours and using linear standardised scores is that it allows clinicians to compare scores on a variety of tests and ultimately across their different distributions. Subsequently, clinicians can then observe extreme scores or outliers in the distribution that may indicate pathology/abnormality or a cognitive strength, depending on which tail it appears However, Mitrushina et al. (2005) are quick to point out that thought needs to be given to other sources of outliers in the distribution. These include, but are not limited to, inadequate reliability; differences in test administration and errors in data

collection; psychosocial, emotional, or motivational effects on the test score; situational factors when testing (e.g., external noise); demographic or physical characteristics of the examinee (e.g., physical handicaps); practice effects; and test biases.

The major limitation of linear standardised scores was emphasised by Nunnally (1978) when he stated, "strictly speaking, test scores are seldom normally distributed" (p. 160). To further emphasise this point, Micceri (1989) conducted an analysis of 440 distributions within the psychometric and psychology research literature with varying populations and settings. He found that no distribution passed all tests of normality. When tests are not normally distributed, the effect is a *skewed distribution*.

2.2.5 Skewed Distributions

Asymmetrical or skewed distributions can either be positive or negative in direction. Figures 2 and 3 depict examples of positive and negative skewed distributions, respectively. A coefficient of skewness can be computed using the following formula.

$$g_1 = \frac{\sum z_i^3}{n} \qquad \text{Formula 8}$$

Where:
z = z score, n = sample size

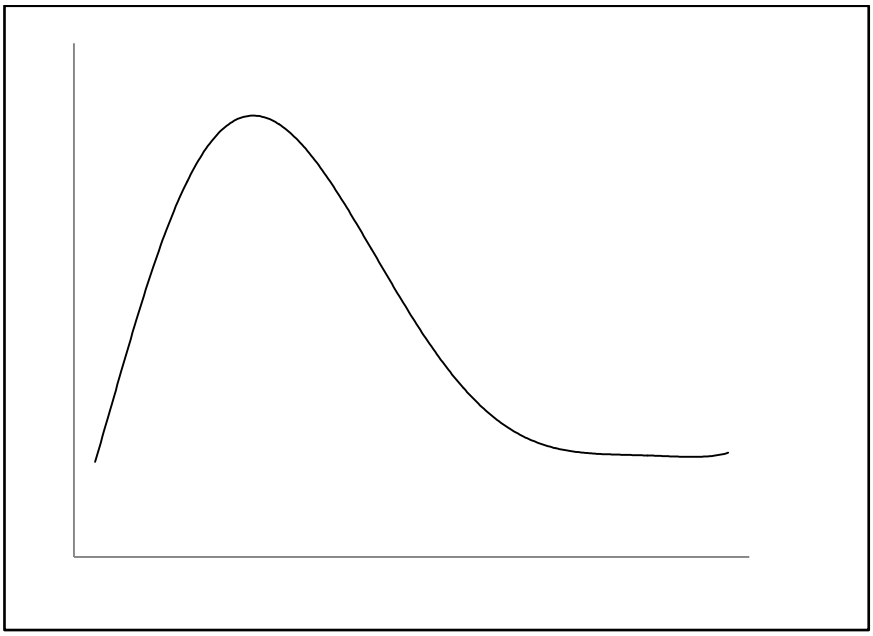

Figure 2.2. Positively skewed distribution

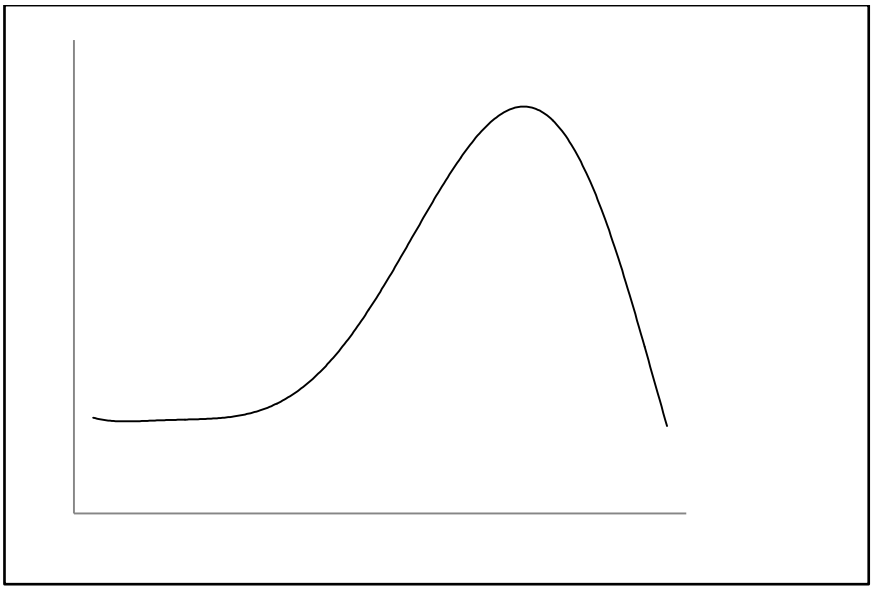

Figure 2.3. Negatively skewed distribution

Negatively skewed distributions have scores that largely fall within the upper half (the median falls above the mean) of the score range and as such have the highest sensitivity at the lower part of the distribution. On the other hand, positively skewed distributions have scores that largely fall within the lower half of the score range (the median falls below the mean) and, therefore, have the highest sensitivity in the upper part of the distribution. Overall, the implication of a skewed distribution is that unlike the normal distribution, the mean and the median are not the same.

The consequence of skewed distributions is that they cannot be readily compared like normal distributions (Anastasi & Urbina, 1997). The mathematical process for comparing normal distributions is well known and can be performed with relative ease. However, it is highly error-prone and extremely difficult to compare skewed distributions in order to make interpretations. Furthermore, linear transformations are unreliable when used on skewed distributions (Crawford et al., 2006). Linear standardised scores are calculated using the normative sample's mean and standard deviation and are based on the assumption of normality, where the median and mean reflect the same measure of central tendency. The problem is that the mean and median of skewed distributions are not the same, and therefore it would be inappropriate to use a linear transformation on a non-normal distribution (Crawford et al., 2006; Mitrushina et al., 2005). The implication, as highlighted by many researchers, is that using the traditional and uncritically accepted linear transformations to interpret raw scores on a skewed distribution will result in multiple errors, including increasing the risk of Type I errors (Brooks et al., 2009; Crawford & Howell, 1998; Crawford et al., 2006). Overall, this may result in over – or under-detection of abnormality, misdiagnosis, unnecessary or incorrect treatment, and/or adverse psychological effects (Strauss, Sherman, & Spreen, 2006). Using a linear transformation on skewed distributions could inaccurately reflect the underlying population rank, skewing the relative standing between the standardised scores.

One solution is forcing skewed distributions into normal distributions (Anastasi & Urbina, 1997). Even though this is distorting the distribution, the power that is gained by doing so is considered to far outweigh the complications and logistical problems of maintaining the skewed distribution.

2.2.6 Normalised Standard Scores

Normalised standard scores have distributions that have been mathematically transformed to fit the normal curve (Anastasi & Urbina, 1997). However, unlike standardised scores, these are mapped on the median, not the mean. Normalised standard scores are created by computing the cumulative frequency of the raw scores of the normative sample and mapping these against the corresponding standard score that represents that percentile on the normal distribution. As such, normalised standard scores can also be referred to as an equipercentile (Budescu, 1987). Clinicians can achieve this by using a percentile conversion table similar to that in Figure 2.4.

Percentile Ranks Corresponding to Different Test Score Scales														
%ile	z	SS	ASS	T	%ile	z	SS	ASS	T	%ile	z	SS	ASS	T
0.03	-3.60	46		14	26	-0.64				77	0.73	111		
0.04	-3.53	47			27	-0.60	91		44	78	0.77			
0.05	-3.47	48		15	28	-0.58				79	0.80	112		58
0.06	-3.40	49		16	29	-0.55				80	0.84			
0.07	-3.33	50	0		30	-0.53	92			81	0.87	113		
0.08	-3.27	51		17	31	-0.50			45	82	0.91			59
0.09	-3.20	52		18	32	-0.47	93			83	0.93	114		
0.1	-3.13	53			33	-0.44				84	1.00	115	13	60
0.2	-3.07	54		19	34	-0.40	94		46	85	1.03			
0.3	-3.00	55	1	20	35	-0.39				86	1.07	116		61
0.4	-2.93	56			36	-0.36				87	1.13	117		
0.5	-2.87	57		21	37	-0.33	95	9		88	1.20	118		62
0.6	-2.80	58		22	38	-0.31			47	89	1.23			
0.7	-2.73	59			39	-0.27	96			90	1.27	119		63
0.8	-2.67	60	2	23	40	-0.25				91	1.33	120	14	
0.9	-2.60	61		24	41	-0.23				92	1.40	121		64
1	-2.53	62			42	-0.20	97		48	93	1.47	122		65
1	-2.47	63		25	43	-0.18				94	1.53	123		
1	-2.40	64		26	44	-0.15				95	1.60	124		66
1	-2.33	65	3		45	-0.13	98			95	1.67	125	15	67
1	-2.27	66		27	46	-0.10			49	96	1.73	126		
1	-2.20	67		28	47	-0.07	99			96	1.80	127		68
2	-2.13	68			48	-0.05				97	1.87	128		69
2	-2.07	69		29	49	-0.02				97	1.93	129		
2	-2.00	70	4	30	50	0.00	100	10	50	98	2.00	130	16	70
3	-1.93	71			51	0.02				98	2.07	131		71
3	-1.87	72		31	52	0.05				98	2.13	132		
4	-1.80	73		32	53	0.07	101			99	2.20	133		72
4	-1.73	74			54	0.10			51	99	2.27	134		73
5	-1.67	75	5	33	55	0.13	102			99	2.33	135	17	
5	-1.60	76		34	56	0.15				99	2.40	136		74
6	-1.53	77			57	0.18				99	2.47	137		75
7	-1.47	78		35	58	0.20	103		52	99	2.53	138		
8	-1.40	79		36	59	0.23				99.1	2.60	139		76
9	-1.33	80	6		60	0.25				99.2	2.67	140	18	77
10	-1.27	81		37	61	0.27	104			99.3	2.73	141		
11	-1.23				62	0.31			53	99.4	2.80	142		78
12	-1.20	82		38	63	0.33	105	11		99.5	2.87	143		79
13	-1.13	83			64	0.36				99.6	2.93	144		
14	-1.07	84		39	65	0.39				99.7	3.00	145	19	80
15	-1.03				66	0.40	106		54	99.8	3.07	146		81
16	-1.00	85	7	40	67	0.44				99.9	3.13	147		
17	-0.95				68	0.47	107			99.91	3.20	148		82
18	-0.93	86			69	0.50			55	99.92	3.27	149		83
19	-0.87	87		41	70	0.53	108			99.93	3.33	150	20	
20	-0.84				71	0.55				99.94	3.40	151		84
21	-0.80	88		42	72	0.58				99.95	3.47	152		85
22	-0.77				73	0.60	109		56	99.96	3.53	153		
23	-0.73	89			74	0.64				99.97	3.60	154		86
24	-0.71			43	75	0.67	110	12		99.98	3.67	155		87
25	-0.67	90	8		76	0.71			57					

Key:	%ile – percentile	SS – Standard Score	ASS – (age) Scaled Score	T – T Score
		Mean = 100 SD = 15	Mean = 10 SD = 3	Mean = 50 SD = 10

Figure 2.4. Conversion Table for Percentiles, z Scores, Scaled Scores, Standard scores, and T-Scores

The corresponding standardised score is normally presented in the same fashion as a linear standardised score, with a mean of zero and standard deviation of one. The following illustrates the normalised transformation of Australian Normative data for the Trail Making Test.

2.2.6.1 Study Two - Normalised Standard Scores for the Trail Making Test

Participants for this study were sourced from a normative database drawn from ongoing normative studies conducted through the Department of Psychology at the University of Southern Queensland (USQ). Participants were generally from regional South-East Queensland or metropolitan Brisbane areas and had volunteered to participate in studies designed to establish Australian norms for a number of neuropsychological tests. In total, 416 cases with Trail Making Test (TMT) data were selected for this study. The TMT is a test designed to measure an individual's "speed for attention, sequencing, mental flexibility and of visual search and motor functioning" (Spreen & Strauss, 1998, p. 533). It has a long and comprehensive history, originally developed for use in the 1944 Army Individual Test Battery (Army Individual Test Battery, 1944). More recent adaptations can be found in several neuropsychological test batteries including the Halstead-Reitan Battery (Reitan & Wolfson, 1993), the Individual Neuropsychological Testing for Neurotoxicity (Singer, 1990), and the Delis-Kaplan Executive Function System (Delis, Kaplan, & Kramer, 2001). The TMT consists of two trials: Trial A (TMT-A) in which participants are required to rapidly draw a line connecting the numbers 1 to 25 in order; and Trial B (TMT-B) in which the participant must alternate between numbers (1 – 13) and letters (A – L) in order.

Of the 416 cases analysed, 240 were female and 176 were male. Ages ranged from 16 to 79 years old (M = 35.97, SD = 15.49) and education from 8 to 20 years (M = 12.93, SD = 2.37). The data were initially stratified into five different age groups, three different education groups, and two gender groups in order to assess the influence of these variables on Part A and Part B test scores. The frequencies of cases in each group are presented in Table 2.8.

Table 2.8.
Number of Cases in Each Stratified Category

Age	*N*	Edu.	*N*	*Gender*	*N*
17 - 20	72	<12	112	Male	176
20-29	101	12	89	Female	240
30-39	86	>12	215		
40-49	64				
>50	92				
Total	416		416		416

The TMT A and TMT B scores were subjected to a three-way analysis of variance (ANOVA) using the three independent variables of age, education, and gender. For TMT A, no main effect was found for gender, $F(1, 386) = 3.19$, $p > .05$, or education, $F(2, 386) = 0.27$, $p > .05$. A significant main effect was found for age, $F(4, 386) = 11.43$, $p < .05$. While there was also a significant age by education effect, $F(8, 386) = 2.58$, $p < .05$, the Tukey *post hoc* test indicated two homogeneous groupings for age; less than 50 years, and more than 50 years, and a single homogeneous group was found for education.

For TMT B, no main effect was found for gender, $F(1, 383) = 0.52$, $p > .05$. However, a significant main effect was found for age $F(4, 383) = 8.24$, $p < .05$ and education $F(2, 383) = 4.14$, $p < .05$. The Tukey *post hoc* test indicated two homogeneous groupings for age, less than 50 years and more than 50 years, and two homogeneous groupings for education, less than 12 years and 12 or more years. As a result, normative data for TMT B was stratified to reflect the 2x2 groupings of age and education. For TMT A, normative data would only need to be stratified by the two levels of age. However, in order to make comparisons between TMT A and TMT B more direct, it was decided that imposing the education structure of the TMT B onto the TMT A, while redundant, would facilitate comparisons between the two trials. The descriptive statistics for the new stratified groups are presented in Table 2.9.

Table 2.9.
Descriptive statistics for mean completion time of TMT A and B for the four groups

Category		TMT-Part A	TMT-Part B
<50 Years Old	*N*	73	72
<12 Years Education	*M*	23.87	60.78
	SD	7.15	23.44
<50 Years Old	*N*	251	250
>12 Years Education	*M*	24.64	54.47
	SD	8.2	19.26
>50 Years Old	*N*	39	38
<12 Years Education	*M*	34.04	79.36
	SD	11.94	29.37
Category		TMT-Part A	TMT-Part B
>50 Years Old	*N*	53	53
>12 Years Education	*M*	29.67	68.44
	SD	9.29	21.69

Based upon the percentiles derived from the cumulative frequencies of the completion times in each group (see Appendix A) scaled scores were allocated to completion time ranges using the percentile range map in Table 2.10.

Table 2.10.
Percentile Rank Ranges Corresponding to Scaled Scores

Scaled Score	Percentile Range
1	≤ 0.5
2	0.6 – 0.9
3	1
4	2 - 3
5	4 - 6
6	7 - 12
7	13 - 20
8	21 - 30
9	31 - 43
10	44 - 56
11	57 - 68
12	69 - 79
13	80 - 87
14	88 – 93
15	94 – 96
16	97 – 98
17	99
18	99.1 - 99.4
19	99.5 +

The results are the corresponding scaled score ranges for the four stratified groups mapped for the TMTA (Table 2.11) and TMT B (Table 2.12).

Table 2.11.
Mapped TMT A Raw Scores to Scaled Scores

	Group			
Scaled Score	<50 years old <12 years ed.	<50 years old 12+ years ed.	50+ years old <12 years ed.	50+ years old 12+ years ed.
1		49.6 +		
2	48.6 +	49.1 – 49.5		
3	42.1 – 48.5	47.6 – 49.0	61.6 +	49.1 +
4	39.1 – 42.0	44.0 – 47.5	60.1 – 61.5	45.1 – 49.0
5	37.6 – 39.0	39.1 – 43.9	51.1 – 60.0	44.1 – 45.0
6	30.1 – 37.5	33.6 – 39.0	48.1 – 51.0	39.1 – 44.0
7	27.1 – 30	30.1 – 33.5	42.6 – 48.0	33.6 – 39.0

8	24.6 – 27	26.6 – 30.0	36.1 – 42.5	31.1 – 33.5
9	23.6 – 24.5	24.1 – 26.5	32.1 – 36.0	29.6 – 31.0
10	21.6 – 23.5	21.6 – 24.0	30.0 – 32.0	27.1 – 29.5
12	18.6 – 19.9	18.1 – 19.5	25.6 – 28.0	23.1 – 24.0
13	17 – 18.5	16.0 – 18.0	21.1 – 25.5	22.1 – 23.0
14	15.6 – 16.9	14.6 – 15.9	18.1 – 21.0	18.1 – 22.0
15	15 – 15.5	13.1 – 14.5	6.1 – 18.0	7.1 – 18.0
16	≤ 14	12.6 – 13.0	≤	≤ 7.0
17		12.1 – 12.5		
18		11.6 -12.0		
19		≤ 11.5		

Table 2.12.
Mapped TMT B Raw Scores to Standard Scores

	Group			
Scaled Score	<50 years old <12 years ed.	<50 years old 12+ years ed.	50+ years old <12 years ed.	50+ years old 12+ years ed.
1		116.1 +		
2	120.1 +	115.1 – 116.0		
3	118.1 – 120.0	111.1 – 115.0		
4	116.1 – 118.0	99.1 – 111.0	140.1 +	124.1 +
5	96.6 – 116.0	79.1 – 99.0	131.0 – 140.0	94.1 – 124.0
6	79.6 – 96.5	70.0 – 79.0	93.1 – 130.0	84.1 – 94.0
7	72.1 – 79.5	64.6 – 69.9	91.0 – 93.0	79.6 – 84.0
8	65.1 – 72.0	59.1 – 64.5	86.6 – 90.0	74.6 – 79.5
10	51.6 – 57.5	49.1 – 53.5	68.1 – 77.5	64.1 – 67.5

11	48.6 – 51.5	45.0 – 49.0	63.1 – 68.0	57.1 – 64.0
12	41.1 – 48.5	40.6 – 44.9	60.1 – 63.0	50.6 – 57.0
13	38.6 – 41.0	36.6 – 40.5	55.1 – 60.0	44.1 – 50.5
14	35.1 – 38.5	30.1 – 36.5	38.1 – 55.0	35.1 – 44.0
15	33.1 – 35.0	27.1 – 30.0	30.1 – 38.0	29.1 – 35.0
16	32.1 – 33.0	24.1 – 27.0	≤ 30.0	≤ 29.0
17	≤ 32.0	23.1 – 24.0		
18		16.1 – 23.0		
19		≤ 16.0		

Overall, this method essentially converts the continuous raw score distribution (time to completion) into a discontinuous scaled score distribution ranging from 1 to 19. The benefit of this is that this data can now be combined with other test scores in a battery to create composite or index scores. To illustrate the use of this system, raw scores for Part A and Part B were standardised using a linear transformation and compared to those standardised using Tables 2.11 and 2.12. For this example, the Part A raw score was 35 seconds and Part B raw scores was 81.5 seconds. The respondent was 35 years old and had 13 years of formal education. For the linear transformation, the *z* scores were calculated using Formula 5 and the descriptive statistics from Table 2.9. Based on these, the *z* score is 1.24 for Part A and 1.40 for Part B. The *z* scores were converted to percentiles using Figure 2.4 with Part A corresponding to the 11^{th} percentile and Part B falling at the 8^{th} percentile. The normalised standard scores corresponding to these raw scores equal a scaled score of 6 (equivalent to 9^{th} percentile) for Part A and a scaled score of 5 (equivalent to 5^{th} percentile) for Part B. As this example demonstrates, using normalised standard scores that are mathematically mapped using the raw score distribution produces scaled scores and corresponding percentiles that are more sensitive than linear transformations.

2.3 Conclusions

When a test is administered, it produces an operational value called a raw score. In order for a clinician to interpret such a score, it needs to be converted into a percentile rank or a standardised score. Both methods have their advantages and disadvantages. For example, while the percentile provides an ordinal position of the client's score and can be readily interpreted by other stakeholders, it fails to provide any information regarding the differences between scores. This type of analysis is particularly important when test batteries are administered. The standardised scores, on the other hand, transform the raw score into a common scale so that analyses can be performed at the battery level. The linear standard score is based around the theoretical normal distribution and as such can introduce error when working with skewed distributions.

What many clinicians may not realise or perhaps choose to ignore, is that skewed distributions are highly exploited in clinical neuropsychology. As such, many neuropsychological tests have distributions that deviate from the normal distribution (Brooks et al., 2009; Capitani & Laiacona, 2000; Crawford & Howell, 1998; Crawford & Garthwaite, 2005). It is important to understand that neuropsychologists and clinicians alike should be particularly interested in skewed distributions because of the predictive and discriminative power they yield (Brooks et al., 2009). The fundamental idea behind cognitive assessments is to identify impairments and strengths. Tests with positively skewed distributions have the majority of their items in the lower end of the distribution and can be particularly effective at differentiating levels of lower performance. Negatively skewed distributions with the bulk of their item content in the upper part of the distribution reflect relatively easy tests for which low scores are generally rare. These often serve well as screening measures as while they do not differentiate levels of poor performance, they are highly sensitive to impairment. Intentionally skewing a test negatively, such as in recognition memory, malingering or effort testing, allows for the highest discriminative power at the lower ability level (Mitrushina et al., 2005). However, another reason why many neuropsychological tests may have skewed distributions is that they often have small *normative sample* sizes and the implications of this is the topic of the next chapter.

CHAPTER THREE
NORMATIVE DATA AND SAMPLE SIZES

3.1 Normative Samples

A normative sample is a group of people who are theoretically representative of the population (Mitrushina et al., 2005). Tests are administered to this normative sample and the scores generated are called *normative data.* Normative data can be collected through a variety of means. One method is that of *census-based norms* where the normative sample is compiled in order to match the demographic variables of a nation's census (Cochran, 1977). The commonly used Wechsler scales utilised this method in creating their extensive normative databases (Wechsler, 1981, 1997, 2008). While this method is thorough and comprehensive, it may disadvantage specific groups. For example, results from a specific ethnic minority group, low education group, or from the extreme elderly may appear as outliers in a normative sample. This may skew the distribution and, more seriously, normal scores from these specific groups may be misinterpreted as impaired (Brooks et al., 2009).

This issue is particularly important in neuropsychology and culturally sensitive testing. For example, developing a normative sample of 100 people in Australia which is intended to be census-matched would be expected to include only two to three indigenous Australians based on the base rate of 2.5 percent of this ethnic group in Australia (Australian Bureau of Statistics, 2011) It is not hard to recognise that this ethnic group would be severely underrepresented by using census-based norming methods. Furthermore, research has found that clinicians are likely to misdiagnose a healthy and cognitively normal African-American as impaired because their scores on cognitive tests are, on average, lower when compared to White American participants (Campbell et al., 2002).

Similar findings have been revealed for low education (Bornstein & Suga, 1988) and the cognitive test performance of normal elderly persons. The study by Marcopulos, McLain, and Giuliano (1997), which sought to generate preliminary normative data for nine common neuropsychological tests, found that many of their participants would be misclassified if the published cut-offs were used. The 133 rural participants in their study were aged over 55 years, had completed no more than 10 years of formal education, and had no history of psychiatric, medical, or neurological disease. The participants were mostly female but there was near equal numbers of White and African Americans. While the conclusions were limited to healthy, low educated, rural, older adults, the study nevertheless highlights the consequences of comparing low frequency groups with the mean for the whole population. Some normative studies, such as the Wechsler scales (The Psychological Corporation, 1997), have attempted to overcome these issues by oversampling some demographic characteristics such as education.

Another method to collect normative data is called the *recruitment method*. This method entails researchers specifying a set of selection criteria and standards and recruiting volunteers based upon these (e.g., healthy volunteers aged 60-70 years with no history of organic or acquired brain damage and with more than 12 years of education). The difference between this method and census-based norms is that the underlying distribution of recruitment norms will not represent the normal population, especially in regard to demographic variables (Williams & Cottle, 1977) but are targeted specifically for the population for which they are intended.

A third method, which is perhaps the most relevant to neuropsychological assessment, is a*nchor norms*. Anchor norms are ideal for neuropsychology because

of the enormous costs and time needed to develop census-based normative samples. Furthermore, because neuropsychological assessment is concerned with testing target populations, demographic variables can be chosen and oversampled in order to meet the needs of the clinician and researcher (Anastasi & Urbina, 1997). Following on from the previous example, while only three indigenous Australians would be sampled for census-based norms, researchers using anchor-norms may choose to sample 100 indigenous Australians to make the sample more appropriate. However, a disadvantage of anchor-norms is the very reason why neuropsychologists and researchers use them: they still tend to have small sample sizes.

3.1.2 Representativeness of the Normative Sample with the Individual

Adequate interpretative comparisons can only be made if the individual being tested is compared with a normative sample that is representative on a variety of levels. Mitrushina et al. (2005) proposed a set of standards to use when selecting appropriate normative samples. They explained:

> "All normative data are limited to use with patients whose demographic characteristics are similar to those of the normative sample and match the administration/scoring procedures of the test utilized (p. 18)."

Without this consideration, it is unknown whether the discrepancies between the individual and the normative sample are reflective of abnormality or the differences between the characteristics of the individual and the normative sample (Ardila, 1995). For example, comparing an 80-year-old individual's response time with the norms of 12-year-olds would probably place the individual in the impaired range, when the difference may only reflect the individual's nonconformity to the characteristics of the normative sample. Overall, this comparison would be prone to errors of both small and large magnitudes. As such, it is important to consider the subject characteristics of the normative data set and the procedures used to obtain them. These commonly include age, gender, education, ethnicity, language, and literacy (Dotson, Kitner-Triolo, Evans, & Zonderman, 2008).

In addition, Mitrushina et al. (2005) highlight the importance of using up-to-date normative samples to take into account changes in the actual test and/or the increases in mean cognitive test performance over time, a phenomenon called the Flynn Effect (Flynn, 1984; 1987; 1994; 1998a; 1998b; 1999). While uncertainty still surrounds the Flynn effect and its implications for psychological testing (Hagan, Drogin, & Guilmette, 2008), the point is nonetheless well-taken that normative samples need to be evaluated for their appropriateness over time particularly with census-based sampling. For example, in 1911 only 3.5% of Australians 18 years or older participated in higher education. By the end of the 20th century in 1996, 65.8% of this same age group were attending an educational institution, (Australian Bureau of Statistics, 2000) highlighting that normative data that is sensitive to the influence of education would need to be re-standardised to accommodate the changes in the population.

Another important standard outlined by Mitrushina et al. (2005) is the notion of sample sizes (*n*). They cited Crawford & Howell (1998) that for normative studies to be deemed adequate and sound, a minimum sample size of 50 is required (Mitrushina et al., 2005). Furthermore, they state "a large number of studies suggest that data based on small sample sizes are highly influenced by individual differences and do not provide a reliable estimate of the population mean" (Mitrushina et al., 2005, p. 70). However, upon analysis of the cell sample sizes found in the studies included in their book, it is apparent that many normative databases do not follow

this rule. For example, for the BNT, there were 28 normative studies included in Mitrushina et al's (2005) book. Out of these, 24 studies had overall sample sizes above 50. However, when analysing the cell sample sizes, only 39 out of 166 cells in the 28 studies had adequate sample sizes of above 50. For JLO, only 11 of the 20 normative studies achieved an overall sample size of 50 or more. For the cell sample sizes, only 17 of the 56 cells had adequate sample sizes. This demonstrates that although there is a standard for sample size in normative studies, the reality is that a large proportion of the published data actually do not conform to this.

3.2 The Optimal Sample Size

Crawford and Howell first mentioned this number, 50, in the literature in 1998 when they developed the Crawford-Howell modified *t*-test to be used on small sample-sizes instead of the conventional *z* score. However, it is important to note that the concept of normative sample sizes needing to have an *n* of 50 is based on an article that actually arrives at a different conclusion. Specifically, Crawford and Howell (1998) suggested "that the modified *t*-test be used with an *n* of less than 50; with larger sample sizes either method is more rapid" (p. 485) where "either method" refers to *z* scores and *t*-tests. In fact, at no point in the article did they suggest that a minimum n of 50 was acceptable for a normative study (J. R. Crawford, personal communication, June, 29, 2013)

More recent research by Bridges and Holler (2007) attempted to determine optimal sample sizes for normative studies. While their study was based on paediatric norms, the results still provide a valid means for questioning and evaluating normative studies in clinical neuropsychology. Their research was two-fold. The first section consisted of calculating confidence intervals and their equivalent *z* scores around the paediatric norms for four common neuropsychological tests: Boston Naming Test (BNT), Rey Auditory Verbal Learning Test (RAVLT), Hooper Visual Organisation Test (HVOT), and the Rey-Osterrieth Complex Figure Test (RCFT). At this initial level, results indicated that the confidence intervals around the normative sample means varied widely, especially when the normative data had a small *n*. For example, on the BNT, normative data for five-year-old boys had a sample size of only 17. When confidence intervals were calculated, the difference between the upper and lower limits was 1.02 standard deviations. Even more unreliable were the 13-year-old girl norms on the same test where the sample size was 4, and the difference between the upper and lower confidence intervals was 3.18 standard deviations.

The second part of the Bridges and Holler (2007) study comprised of recalculating the confidence intervals for the same paediatric normative studies, but with different sample sizes (i.e., n = 5, 10, 25, 50, 100, 200, 300, and 500). Bridges and Holler (2007) concluded:

> "Fewer than 50 subjects results in confidence intervals that are deemed too large to be of clinical utility to neuropsychologists. Alternatively, normative studies having more than 75 subjects per group may not significantly decrease the width of the confidence interval" (p. 537).

One limitation reported by the researchers is that their study focused primarily on the effect sample size had on normative sample means. They highlighted that sample standard deviations are also affected by sample size. They also pointed out their study calculated confidence intervals around normative means that were based on a normal distribution (Bridges & Holler, 2007). Therefore, for

highly skewed distributions, their optimal sample size recommendations may not be applicable.

Crawford and Garthwaite (2008) disagreed with the optimal sample size recommendations of Bridges and Holler (2007). They stated "the 'one size fits all' approach is inappropriate" (p. 112). They expressed the view that clinicians can benefit from small sample sizes, but that "every effort should be expended to make the sample as large as practical constraints allow" (Crawford & Garthwaite, 2008, p. 112).

It is interesting to consider that some researchers have approached this same problem of adequate sample sizes but with regard to creating stable reliability coefficients. Nunnally (1978) specified that 300 participants were required for a sample to create a stable reliability coefficient. Charter (1999) concluded that a minimum of 400 participants was needed for accurate and stable estimates of reliability for an individual test score. He later stated "the larger the *n* the greater the precision there is in estimating the population reliability coefficient (Charter, 2001, p. 693). This suggests that even if smaller sample sizes were capable of generating stable measures of central tendency and variance, much larger sample sizes may be needed to generate a comparable level of stability in another psychometric characteristic.

It is an interesting aside that clinicians reveal important underlying assumptions in the way and extent to which they apply psychometric properties to their clinical decision-making. While virtually all practitioners are aware of reliability and its influence on test score error, and may generate confidence intervals to express the scope of that error, they nonetheless treat the reliability coefficient for the test as if it is invariant and itself has no error. One wonders if clinicians would care to compute confidence intervals if the upper and lower bounds of the score had to be computed based on the upper and lower bounds of the reliability coefficient.

3.2.1 Neuropsychology and Sample Sizes

Neuropsychology is particularly vulnerable to the issue of sample size because of the nature of the discipline. While some studies use optimal sample sizes, most fail to recognise that many tests are stratified by demographic variables (Crawford & Howell, 1998). For example, Forrester and Geffen's **(**1991**)** Australian children's norms for the Rey Auditory Verbal Learning Test (RAVLT) have an ostensibly adequate sample size of 80, exceeding the Bridges & Holler (2007) suggestion of 75. However, once these norms are stratified by age (7-8; 9-10; 11-12; and 14-15) and gender, the resulting *n* for each cell is only 10. If children assessed with the RAVLT were intended to be compared to the grand mean and standard deviation of the study this may be an acceptable sample size. However, the stratification of the normative data was based upon the determination that both age and gender influence performance on the RAVLT. Applying the minimum n of 75 to each stratified cell would suggest a minimum sample size of 600. For this purpose the sample size was woefully inadequate. Similarly, the Australian child normative data study of six tests including the RAVLT collected by Anderson and Lajoie in 1996 suffers from the same drawback. While they report a large sample size of 376, they also stratify their data by gender and seven age groups. As a result, they have sample sizes ranging from 18 to 33. It is interesting to note that these two Australian normative studies published in the early to mid 1990's still constitute the main published norms for children and adolescents for this test.

Another reason for small sample sizes is that most normative studies require a great deal of time and resources to obtain large samples (Crawford & Garthwaite, 2002). Unlike other disciplines of psychology, development of normative data is not only conducted by test publishers, but is also performed by independent clinicians who benefit from more specific anchor-normed studies. This is particularly apparent for neuropsychological tests that were developed without affixed norms (Williams & Cottle, 2011). Forrester and Geffen's (1991) norms are a prime example of this. The RAVLT was initially popularised in North America in the first edition of *Neuropsychological Assessment* (Lezak, 1976) and provided normative data from Swiss and French samples of Labourers (N = 25), Professionals (N = 30), Students (N = 47), Elderly Labourers (N = 15), and Elderly Professionals (N = 15) (Rey, 1964). The onus has been, and still is, on clinicians to develop more appropriate norms that would benefit their own clinical practice. Another reason is that many neuropsychological tests do not have clear copyright statuses (Williams & Cottle, 2011) and test publishers may prefer not to invest in large-scale norming studies because of the commercial risks involved. For example, the copyright status of the RAVLT is uncertain given that it was developed more than 50 years ago using a list of words that was originally developed by Édouard Claparède for his "Test de mémoire des mots", a single trial memory test developed between 1916 and 1919 at the University of Geneva (Boake, 2000)

3.2.2 Meta-norming

Meta-norming is the process of combining a variety of individual studies through regression analyses in order to develop collective normative data sets for particular cognitive tests (Mitrushina et al., 2005). Analyses take into account demographic variables, sample sizes, and the version of the tests used, administration procedures, recency of the studies, recruitment strategies, scoring procedures, and reporting of IQ levels. Overall, meta-norming is intended to allow clinicians to make clinical interpretations based on large compiled normative database for any particular test.

While there are some significant advantages to meta-norming, particularly with regard to increasing normative sample sizes, there are two major limitations: error from the underlying normative samples, and the presence of recruitment bias. The former is related to the representativeness of the normative sample. The process of meta-norming may introduce error from the level of disparities between the meta-norming of normative samples (Mitrushina et al., 2005). For example, it is highly improbable that adolescent norms from one study have the same characteristics as the adolescent norms of another on the same test. This type of error, however, is evident regardless of whether a clinician utilises meta-norms. That is, any clinical interpretations based on a variety of tests will utilise normative data that are dissimilar in their demographic characteristics and testing procedures (Russell, Russell & Hill, 2005). While Mitrushina et al. commented on these issues, they outlined selection criteria for their large-scale meta-norming study. For example, studies on a particular cognitive test are only included if they examine the same version of the test and have the same administration procedures. As such, research that failed to include standardised administration procedures and demographic details about their samples, had idiosyncratic samples, or did not provide sample statistics were not included in the meta-norming process (Mitrushina et al., 2005)

Recruitment bias is the second limitation of meta-norming. A study by Williams and Cottle (2011) compared census-based norms of the WAIS-R with their

own meta-analysis of independent norms published on the WAIS-R from 1981 to 2009. The independent norms were combined using the procedures outlined by Mitrushina et al. (2005). Comparisons found levels of recruitment bias within the meta-norming samples. The meta-analysis sample had an average of two more years of education than the published WAIS-R norms. There were also more women in the community-based normative data within the meta-analysis sample than in the WAIS-R norms. Race distribution was also different between the two sets of normative data, with WAIS-R norms including more non-white-identified subjects (13% compared to 7% in the meta-analysis). When the meta-analysis norms were corrected for this recruitment bias, using a Cholesky decomposition (Mooney, 1997) and adjusting the scores using the Deming method (Deming & Stephan, 1940) to match the census-based norms used with the WAIS-R, results indicated that the summary statistics were comparable with the WAIS-R published norms.

Meta-norming is particularly useful for clinicians who adopt the "flexible" or semi-flexible approach to a cognitive test battery. However, in a "fixed battery" this approach is not warranted because of their comprehensive and co-normed construction.

3.2.3 Co-norming in "Fixed" Batteries

Batteries that are "fixed" have the advantage of being co-normed. Each test in the battery is normed together using the same procedure and normative sample (Russell et al., 2005). This approach, therefore, eliminates the discrepancies that are found when different normative samples for several neuropsychological tests are used. Furthermore, when a variety of tests is standardised together they can be corrected for the probabilities of scores obtained in the impaired range when no impairment exists (Russell et al., 2005). Overall, a "fixed" battery can provide validated cut-scores that help the clinician to determine which scores are in the range associated with impairment and whether the impairment is significant and/or abnormal.

Overall, the fact remains that the practitioners are still divided and perhaps unaware of the optimal *n* that is needed in order to obtain stable means and standard deviations let alone the issue raised by Bridges and Holler (2007) of non-normal or skewed distributions. It is, therefore, important to evaluate the sample size needed to get stable means and standard deviations for different skewed distributions of tests commonly employed in neuropsychology. No study has evaluated the clinical consequence of having data cells with inadequate sample sizes and this is the basis for the following study.

3.3 Finding the Optimal N – Study Two

Study Two is concerned with finding the optimal *n* in order to produce stable means and standard deviations from different distributions.

Participants:

Participants for this study were sourced from three separate databases. One normative sample used for analysis was drawn from ongoing normative studies conducted by the Department of Psychology at the University of Southern Queensland (USQ). Participants were generally from regional South-East Queensland or metropolitan Brisbane areas and had volunteered to participate in studies designed to establish Australian norms for a number of neuropsychological tests.

Two additional databases were used for Study Two. The standardisation and educational oversampling normative data for Wechsler Adult Intelligence Scale – Third Edition (WAIS-III; Wechsler, 1997a) Symbol Search was sourced from Lange, Chelune, Taylor, Woodward, & Heaton (2006). Sampling characteristics for this data are described in the WAIS-III/WMS-III Technical Manual (The Psychology Corporation, 1997). In addition, data for the Rey 15 Item test was from an archival clinical database of personal injury litigant cases assessed in a forensic psychological practice in Brisbane, Australia.

Materials:

Due to differences in testing protocols over time only protocols that contained data for Trail Making Test A and B (TMT A, TMT B), Controlled Oral Word Association Test (COWAT), WAIS-III Symbol Search (WAIS-III SS), Wechsler Test of Adult Reading (WTAR), Rey 15 Item Test (Rey 15 Item) or Hooper Visual Organisation Test (HVOT) were analysed. The COWAT is a verbal fluency test which requires the test-taker to spontaneously name words beginning with the letters F, A, and S. The version analysed is an update from the original test developed by Benton, Hamsher, and Sivan (1983). WAIS-III Symbol Search (Wechsler, 1997a) is a visual scanning and processing subtest on the WAIS-III used to assess a test-taker's ability to detect the presence of one or more target symbols in a sequence of five. The WTAR (Wechsler, 2001) is a reading test that includes 50 of irregularly pronounced words ordered in increasing difficulty and was developed using methodology directly associated with the NART. It was developed and co-normed with the WAIS-III (Wechsler, 1997a) and the Wechsler Memory Scale – Third Edition (WMS-III; Wechsler, 1997b) and can be used as an estimate of premorbid functioning. The Rey Fifteen Item Test (Lezak, 1995) is a brief memory test used to detect inadequate cognitive effort. Lastly, the HVOT (Hooper, 1952; Western Psychological Services, 1983) is a visual perception and discrimination test which requires test takers to identify common objects and animals that are cut-up and rearranged in an unsystematic way. It should be noted that there are new editions for many of the tests used in the current study. These tests were ultimately selected because of the large sample sizes they possess and because they served as a means of illustrating the methodology.

These tests were chosen because of the hypothesised differences in skewness underlying their distributions. The Statistical Package for the Social Sciences (SPSS) version 21 was used for the analyses. Pairwise deletion was used to separate protocols that did not contain data for these tests. Table 3.1 presents the sample sizes for each of the seven chosen measures and the database from which they were acquired.

Table 3.1.
Sample Sizes and Databases for Seven Neuropsychological Tests

Tests	Sample Size (*N*)	Database
TMT A	507	USQ Normative
TMT B	507	USQ Normative
COWAT	935	USQ Normative
WAIS-III Symbol Search	1250	WAIS-III Standardisation
WTAR	389	USQ Normative
Rey 15 Item	272	Clinical Data
HVOT	379	USQ Normative

Note: COWAT = Controlled Oral Word Association Test; HVOT = Hooper Visual Organisation Test; TMT A = Trail Making Test Subtest A; TMT B = Trail Making Test Subtest B; WAIS-III SS = WAIS-III Symbol Search; WTAR = Wechsler Test of Adult Reading

The demographic characteristics for each of the tests were calculated and are presented in Table 3.2.

Table 3.2.
Descriptive Statistics for the Seven Tests

				Age	Education	Gender		
Test	*M*	*SD*	*Mdn*	*M* (*SD*)	*M* (*SD*)	M	% Males	F
TMT A	25.95	9.20	24	35.66 (15.17)	13.10 (2.37)	212	42%	295
TMT B	59.61	22.69	55	35.66 (15.17)	13.10 (2.37)	212	42%	295
COWAT	42.11	11.70	41	36.04 (14.17)	12.78 (2.30)	383	41%	552
WAIS-III SS	28.22	10.50	29	48.36 (23.96)	12.47 (2.60)	581	46%	669
WTAR	37.32	8.05	39	40.19 (14.12)	12.82 (2.50)	168	43%	221
Rey 15 Item	13.02	2.67	15	37.73 (12.82)	11.56 (2.60)	187	69%	85
HVOT	26.43	2.55	27	35.49 (16.05)	12.5 (2.40)	150	38%	229

All test distributions were then analysed for skewness, and normal area curve histograms created for each measure. Non-normality of the distributions was determined using Kolmogorov-Smirnov statistics and rudimentary skewness classifications developed by Bulmer (1979). The Kolmogorov-Smirnov statistics assess normality of a distribution of scores. Table 5 presents normality information for each test and are characterised using Bulmer's classifications as follows:

- Distribution is approximately normal if skewness is between -0. 5 and + 0.5
- Distribution moderately skewed if skewness is either between -1 and -0.5 or + 1 and + 0.5
- Distribution is extremely skewed if skewness is either less than -1 or greater than + 1

Table 3.3.
Tests of Normality for the Seven Tests

					Skewness Statistics	
Test	Skewness	*SEE*	Bulmer	SE_E	Statistic	Sig.
TMT B	1.70	0.11	Extremely	0.22	0.12	0.00
TMT A	1.40	0.11	Extremely	0.22	0.12	0.00
COWAT	0.60	0.08	Moderately	0.16	0.05	0.00
WAIS-III SS	0.00	0.70	Normal	0.14	0.05	0.00
WTAR	-0.81	0.12	Moderately	0.25	0.09	0.00
Rey 15 Item	-1.42	0.15	Extremely	0.29	0.29	0.00
HVOT	-1.64	0.13	Extremely	0.25	0.12	0.00

Note: Bulmer refers to Bulmer's (1979) classifications. Skewness Statistics = Kolmogorov-Smirnov Statistics. *SEE* = standard error of the estimate.

The seven neuropsychological tests chosen for this study range in degree of skewness. TMT A and TMT B are both extremely positively skewed whereas Rey 15 Item Test and HVOT are extremely negatively skewed. The Kolmogorov-Smirnov test statistics for each of these tests also indicate non-normality ($p< 0.05$). A moderately positively skewed test (COWAT) and a moderately negatively skewed test (WTAR) were also included. What is interesting to note is that the WAIS-III Symbol Search's skewness statistic would be classified as "Normal" according to Bulmer (1979), but is non-normal according to the Kolmogorov-Smirnov test. Pallant (2010), however, explains that the Kolmogorov-Smirnov test has high power in large samples, and that it is quite common to get significant values. In other words, the Kolmogorov-Smirnov test is suggesting that the WAIS-III Symbol Search distribution differs significantly from a normal population, even though the said deviation is not large enough to cause an issue with the skewness statistic that

assumes normality. Given this, the WAIS-III Symbol Search test will be considered "normally distributed". It is important to note here that in social sciences a distribution with skewness and kurtosis equalling zero is quite uncommon (Pallant, 2010).

Distributions for each test were transformed into histograms. These are presented below in Figures 3.1 through 3.7. Distribution curves have been incorporated into each histogram as a visual indicator of degree of skewness and kurtosis. Also included are lines indicating the mean and median of each distribution.

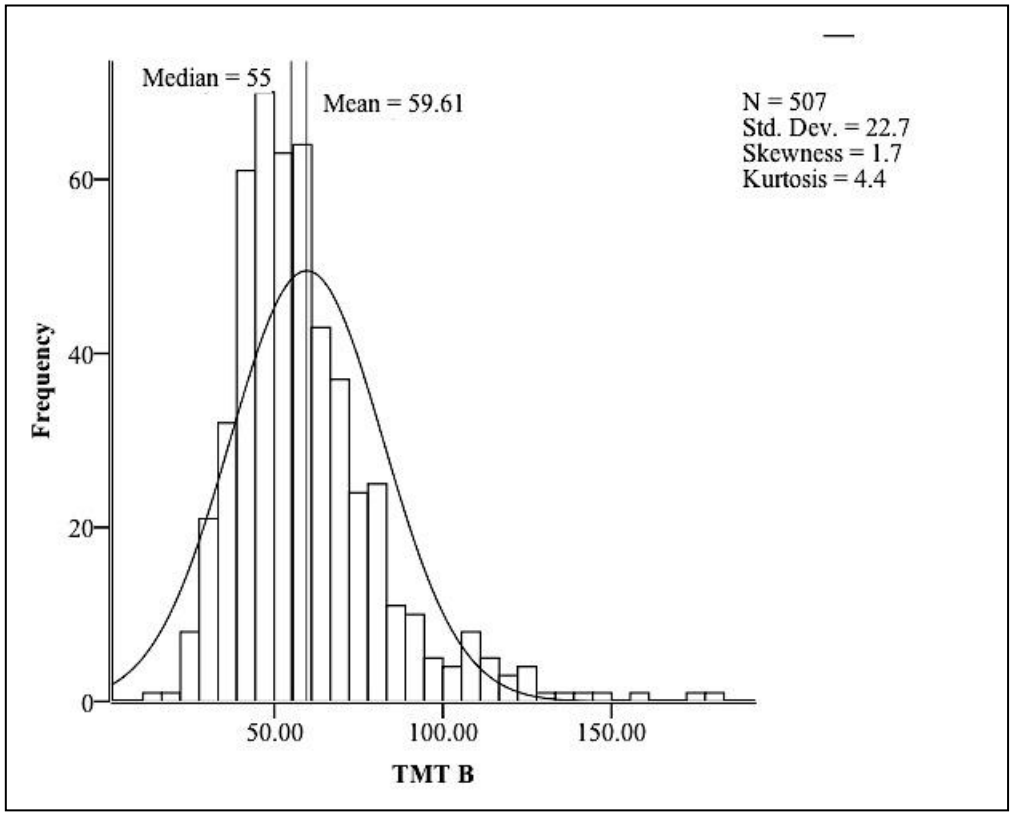

Figure 3.1. Histogram of TMT B distribution

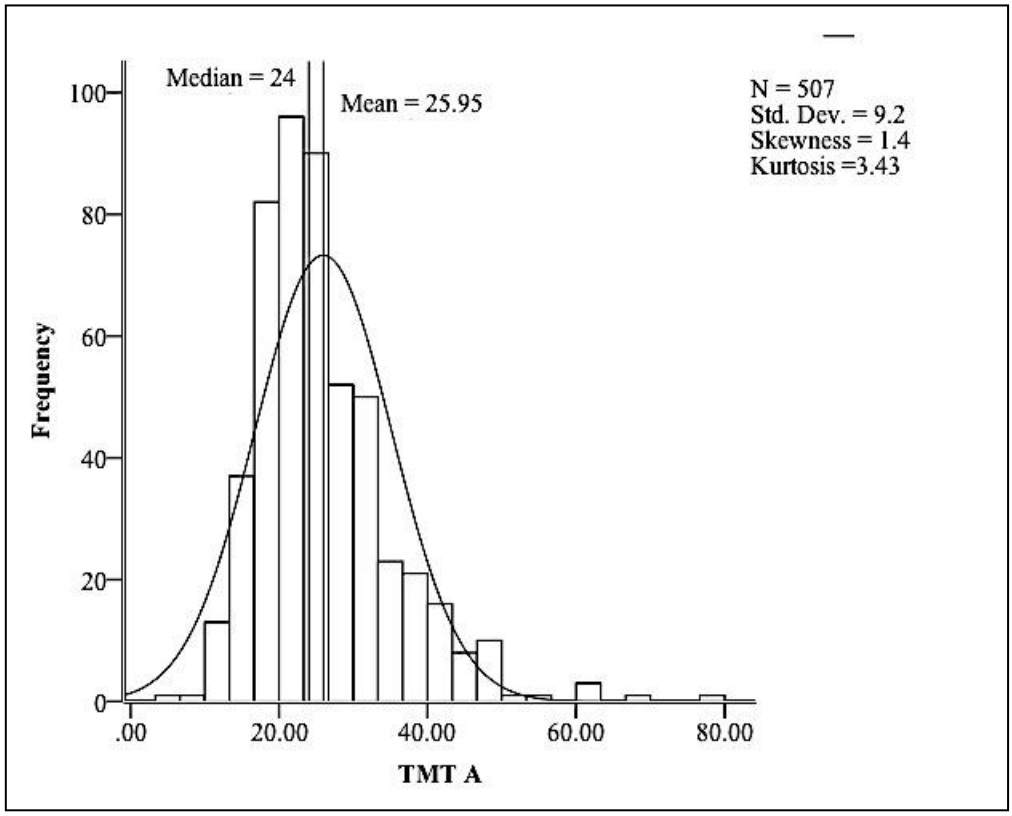

Figure 3.2. Histogram of TMT A distribution

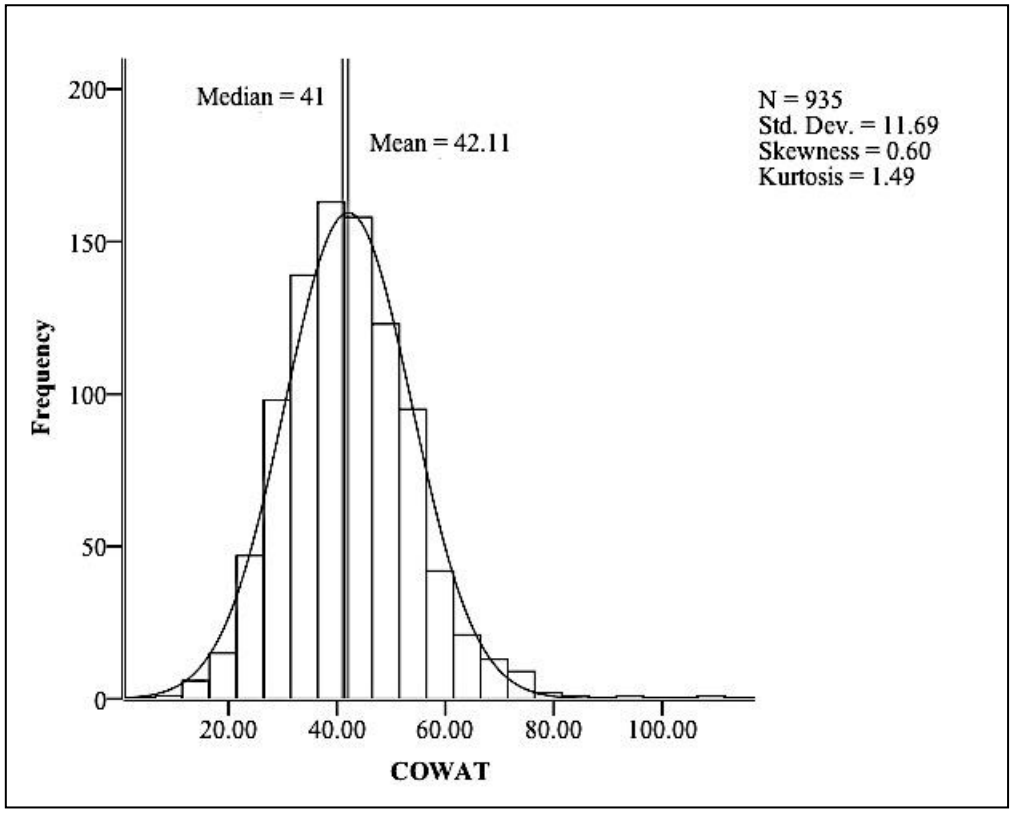

Figure 3.3. Histogram of COWAT distribution

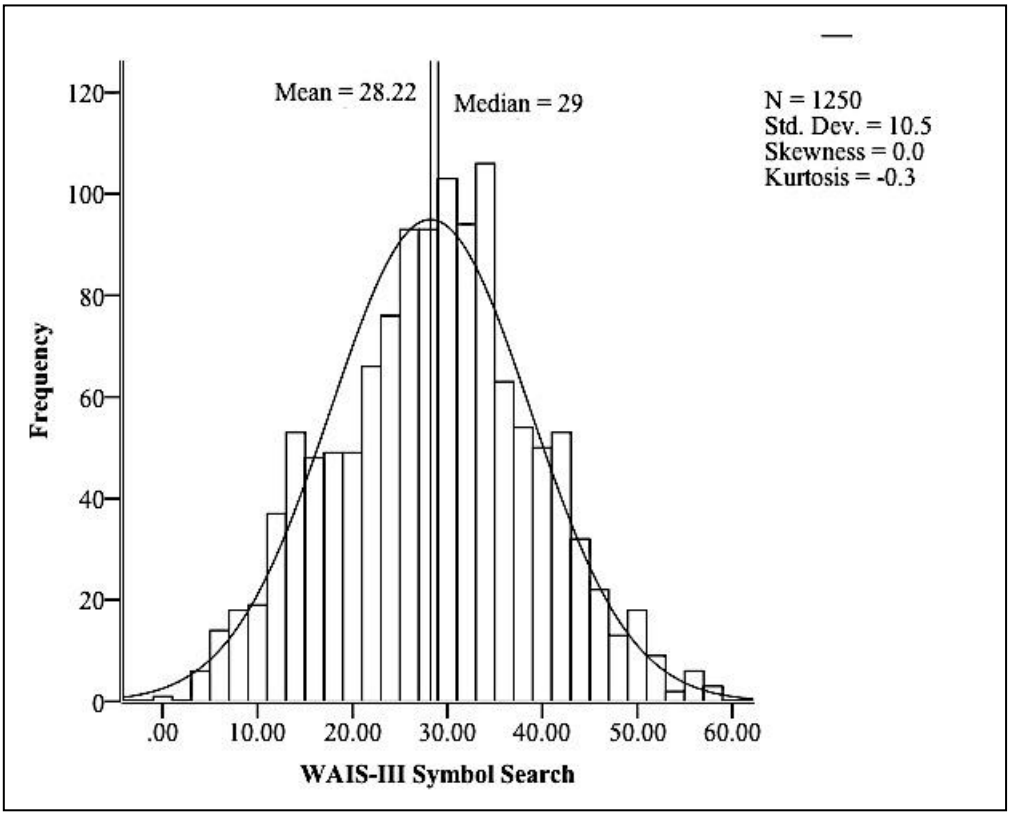

Figure 3.4. Histogram of WAIS-III Symbol Search distribution

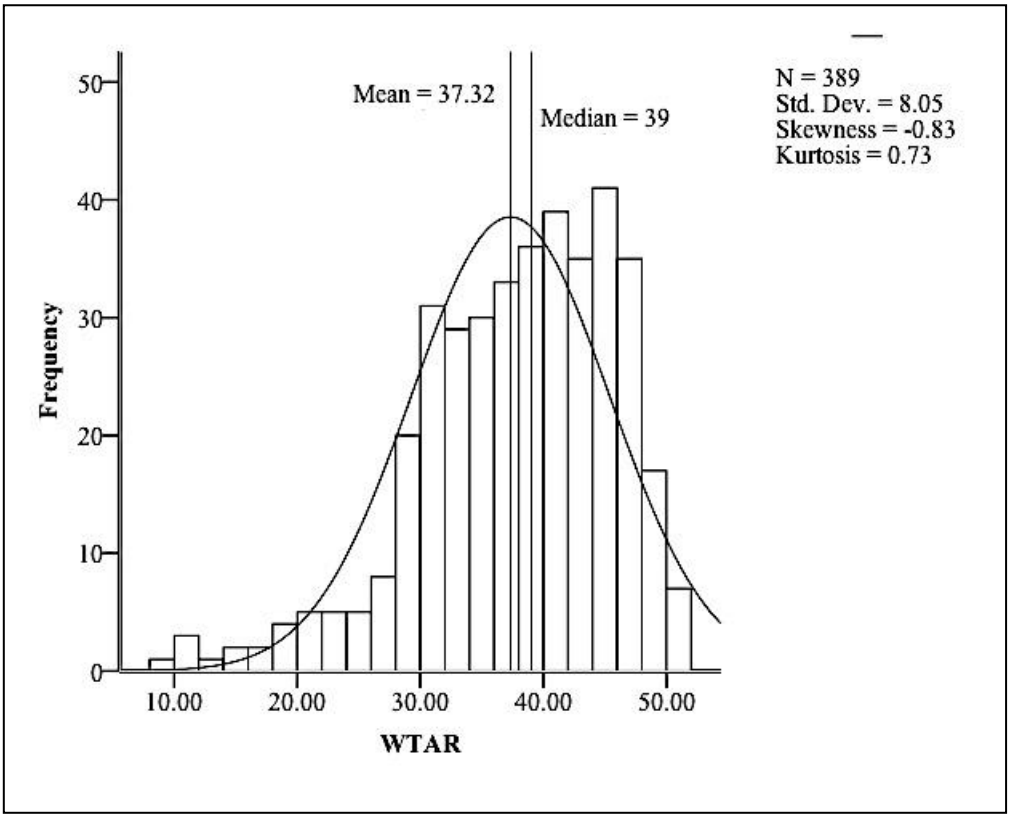

Figure 3.5. Histogram of WTAR distribution

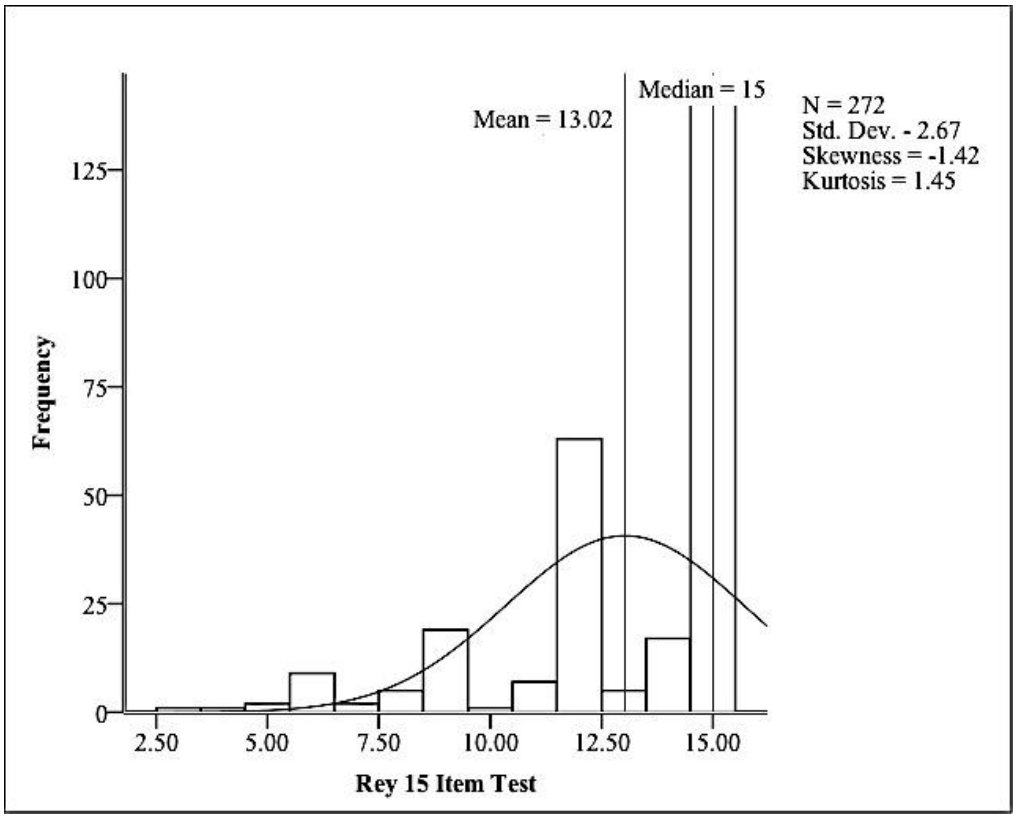

Figure 3.6. Histogram of Rey 15 Item Test distribution

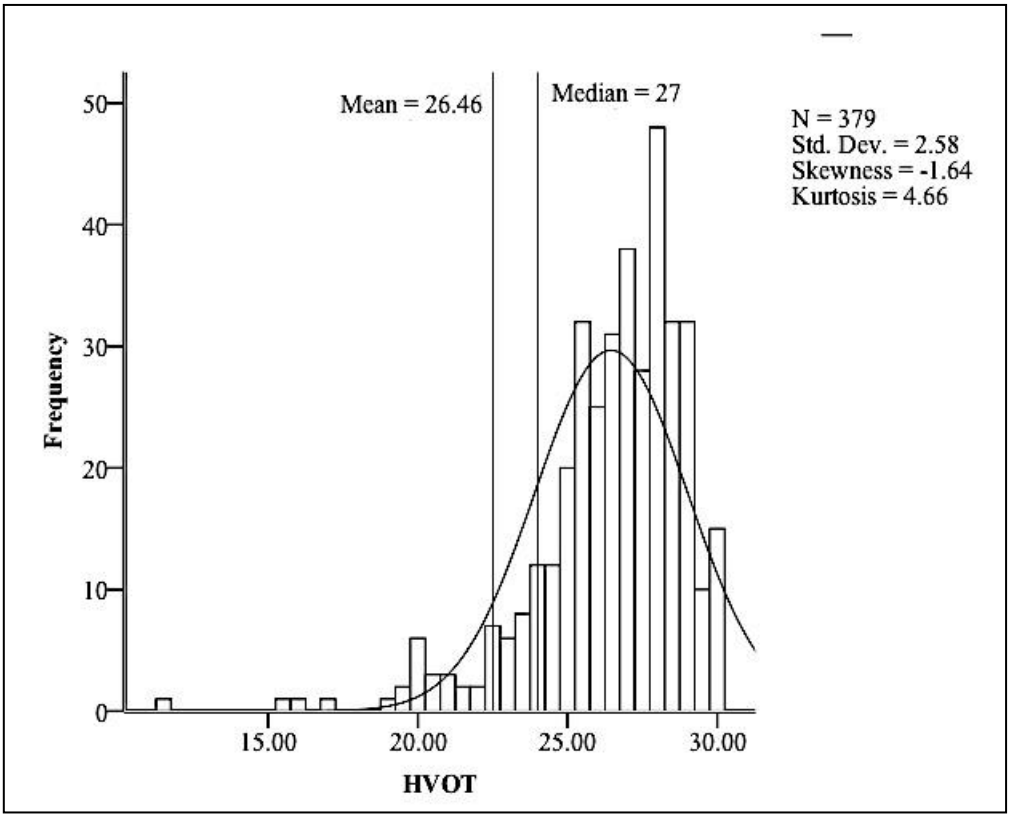

Figure 3.7. Histogram of HVOT distribution

SPSS 21 was then utilised to randomly draw cases with replacement from each of the seven tests, producing sample sizes of 10, 20, 30, 40, 50, 60, 70, 80, 90,

and 100. For TMT A, TMT B, and COWAT, the sample sizes also included 110 and 120. For each sample size, the process was completed five times and the average mean and variances were calculated. Technically, only one sample should be drawn for each sample size as this replicates the method used with small *n* normative studies. However, multiple samples are required to generate a variance measure for the stnadard deviations, necessitating multiple samples to compute an average standard deviation. This was achieved by drawing of five samples with replacement for each sample size. This number was chosen in an effort to keep the sampling as small as possible, consistent with the spirit of the thesis.

Average standard deviations were computed by averaging the variances and taking the square root of the resulting result. Descriptive statistics for each test and the sample parameters for each sample size are displayed in Tables 3.4 through 3.10. It should be noted that the mean (μ) and standard deviation (σ) of each distribution are now treated as population statistics, whereas the means and standard deviations for the samples are treated as sample statistics.

Table 3.4.

Descriptive Statistics for TMT B as a Function of *Sample Size* ($\mu = 59.61$; $\sigma = 22.70$)

n	10	20	30	40	50	60	70	80	90	100	110	120
Sample *M*	62.47	63.78	57.57	55.11	61.67	59.43	59.10	59.25	59.63	59.84	59.63	59.94
Sample *SD*	20.71	24.94	20.85	17.65	22.34	23.40	21.86	21.27	21.38	22.30	21.85	23.07

Table 3.5.

Descriptive Statistics for TMT A as a Function of *Sample Size* ($\mu = 25.95$; $\sigma = 9.20$)

n	10	20	30	40	50	60	70	80	90	100	110	120
Sample *M*	26.22	26.57	24.96	24.93	26.5	25.92	25.18	25.87	25.98	26.35	26.01	25.91
Sample *SD*	7.01	9.59	8.98	9.30	9.26	9.34	8.94	9.05	9.36	9.32	9.37	9.51

Table 3.6.

Descriptive Statistics for COWAT as a Function of *Sample Size* ($\mu = 42.11$; $\sigma = 11.70$)

n	10	20	30	40	50	60	70	80	90	100	110	120
Sample *M*	43.32	41.25	41.66	41.67	42.24	40.61	40.74	41.85	42.23	42.12	41.96	41.97
Sample *SD*	10.40	10.81	10.32	12.02	11.31	10.92	10.93	11.20	11.20	11.20	11.19	11.53

Table 3.7.

Descriptive Statistics for WAIS-III Symbol Search as a Function of *Sample Size* ($\mu = 28.22$; $\sigma = 10.50$)

n	10	20	30	40	50	60	70	80	90	100
Sample *M*	28.28	29.06	28.55	27.53	27.09	28.22	28.14	28.12	28.06	28.39
Sample *SD*	11.68	11.06	10.28	9.67	10.01	9.98	10.58	10.72	10.29	10.71

Table 3.8.

Descriptive Statistics for WTAR as a Function of *Sample Size* ($\mu = 37.32$; $\sigma = 8.05$)

n	10	20	30	40	50	60	70	80	90	100
Sample *M*	37.32	37.01	36.98	38.70	38.16	37.54	37.41	37.06	37.05	37.47
Sample *SD*	6.73	6.92	8.70	7.92	8.29	8.03	7.84	8.47	8.36	7.98

Table 3.9.

Descriptive Statistics for Rey 15 Item as a Function of *Sample Size* ($\mu = 13.02$; $\sigma = 2.67$)

n	10	20	30	40	50	60	70	80	90	100
Sample *M*	13.40	13.14	13.20	13.09	13.16	13.04	12.95	12.9	13.05	13.10
Sample *SD*	2.40	2.48	2.51	2.47	2.54	2.51	2.76	2.82	2.62	2.55

Table 3.10.

Descriptive Statistics for HVOT as a Function of *Sample Size* ($\mu = 26.46$; $\sigma = 2.55$)

n	10	20	30	40	50	60	70	80	90	100
Sample *M*	26.32	26.13	26.63	26.53	26.59	26.44	26.45	26.59	25.35	26.35
Sample *SD*	2.32	2.20	2.45	2.38	2.38	2.78	2.51	2.41	2.60	2.52

To account for sampling error, 90% confidence intervals were calculated for the population statistics for each test. For the population mean, the standard error of the mean was calculated using Formula 9 below. This was then utilised in Formula 10 to calculate upper and lower 90% confidence intervals.

$$SEM = \frac{\sigma}{\sqrt{n}} \qquad \text{Formula 9}$$

Where:

σ = *the population standard deviation*
n = the sample size

$$\mu \pm (SE \times 1.645) \qquad \text{Formula 10}$$

Where:

μ = *the population mean*
SEM = the standard error of mean

Ninety percent confidence intervals for the population standard deviation were calculated using Bootstrap, a computer program based on the general assumption that the characteristics of a population can be approximately determined by a random sample of that sample population (Field, 2009). It is commonly applied to determine confidence intervals of population parameters or statistics (Sheskin, 2004). For this study, Bootstrap was set to obtain *m* = 1000 bootstrap samples (where *m* is subsamples) and calculate the 90% confidence interval around the estimated population standard deviation. This technique was applied to all seven tests. Table 3.11 below presents the 90% confidence intervals for the population statistics for each test and the standard error of means.

Table 3.11.
90% Confidence Intervals for Population Means and Standard Deviations

		90% Confidence Intervals			
		M		*SD*	
Test	*SEM*	Upper	Lower	Upper	Lower
TMT B	1.01	61.27	57.95	24.83	20.61
TMT A	0.41	26.63	25.28	10.00	8.45
COWAT	0.38	42.74	41.49	12.29	11.07
WAIS-III SS	0.30	28.71	27.73	10.84	10.2
WTAR	0.41	38.00	36.65	8.57	7.48
Rey 15 Item	0.16	13.28	12.76	2.90	2.40
HVOT	0.13	26.67	26.25	2.80	2.28

These confidence intervals were then applied to the sample means and standard deviations of each sample size. Figures 3.8 through 3.20 display graphs of the sample means and standard deviations with corresponding 90% confidence intervals for each of the seven measures. For each figure, the point at which the mean and the standard deviations stabilise is defined as the point at which the sample means and standard deviations consistently fall within the 90% confidence intervals and indicates the optimal sample size.

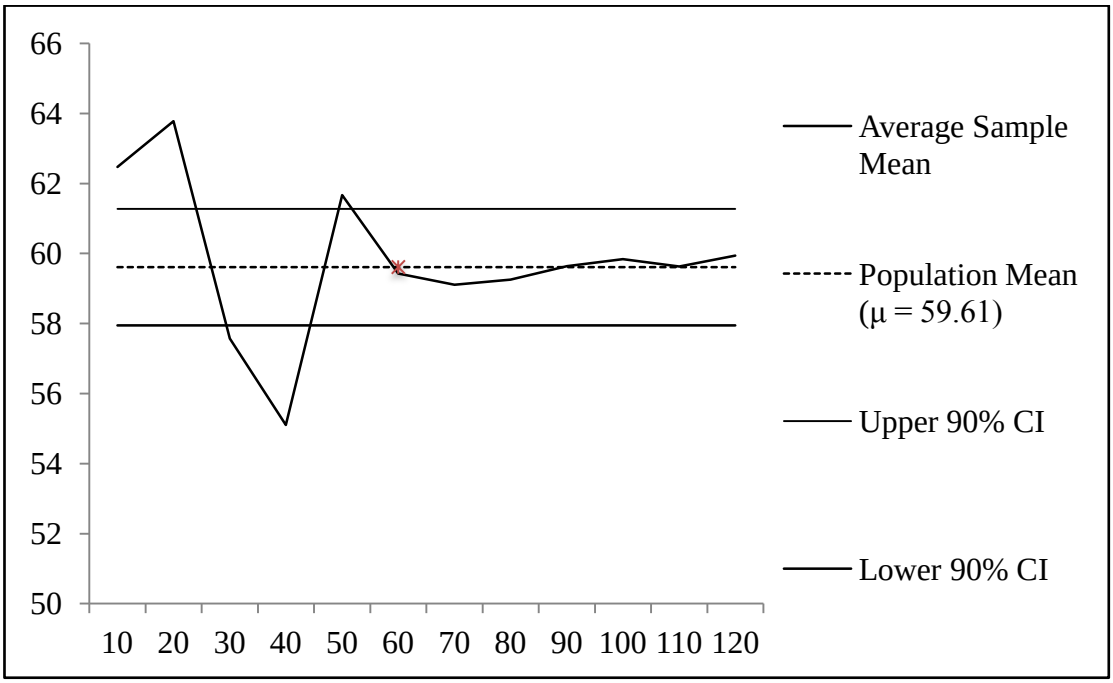

Figure 3.8. TMT B sample means with 90% confidence intervals

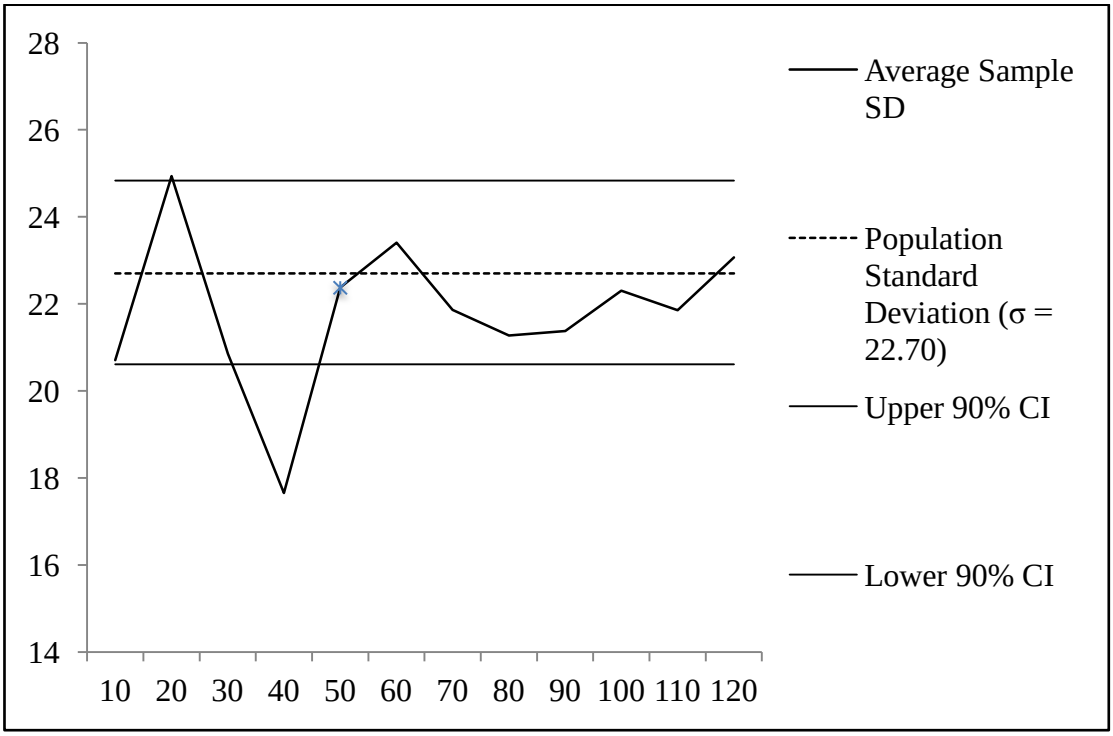

Figure 3.9. TMT B sample standard deviations with 90% confidence intervals

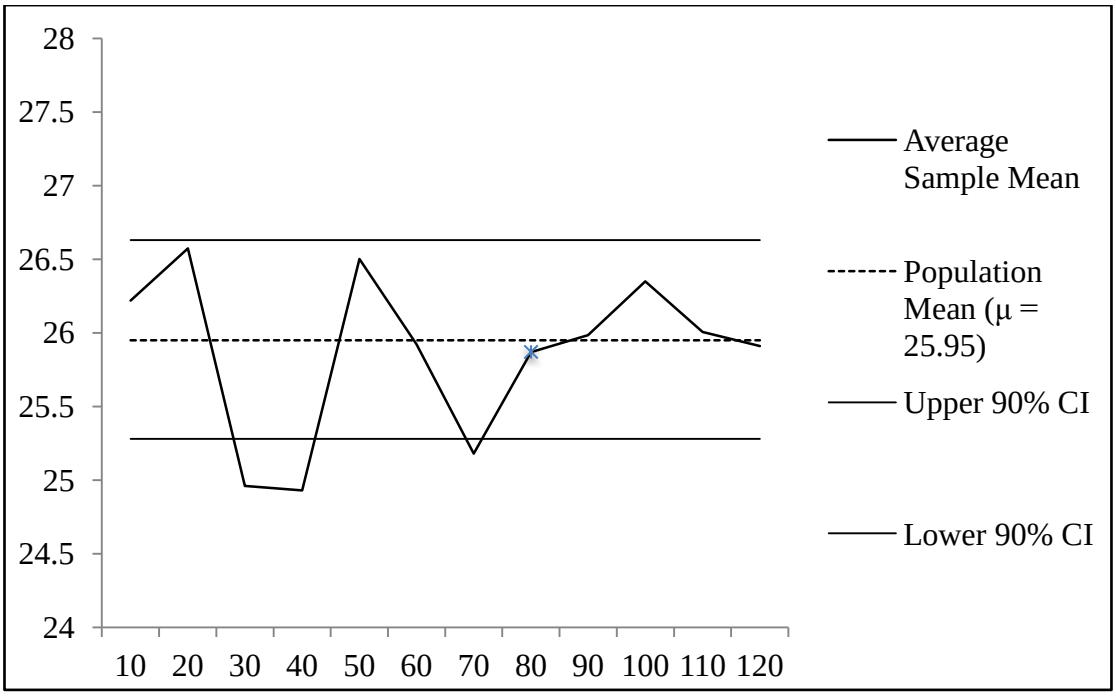

Figure 3.10. TMT A sample means with 90% confidence intervals

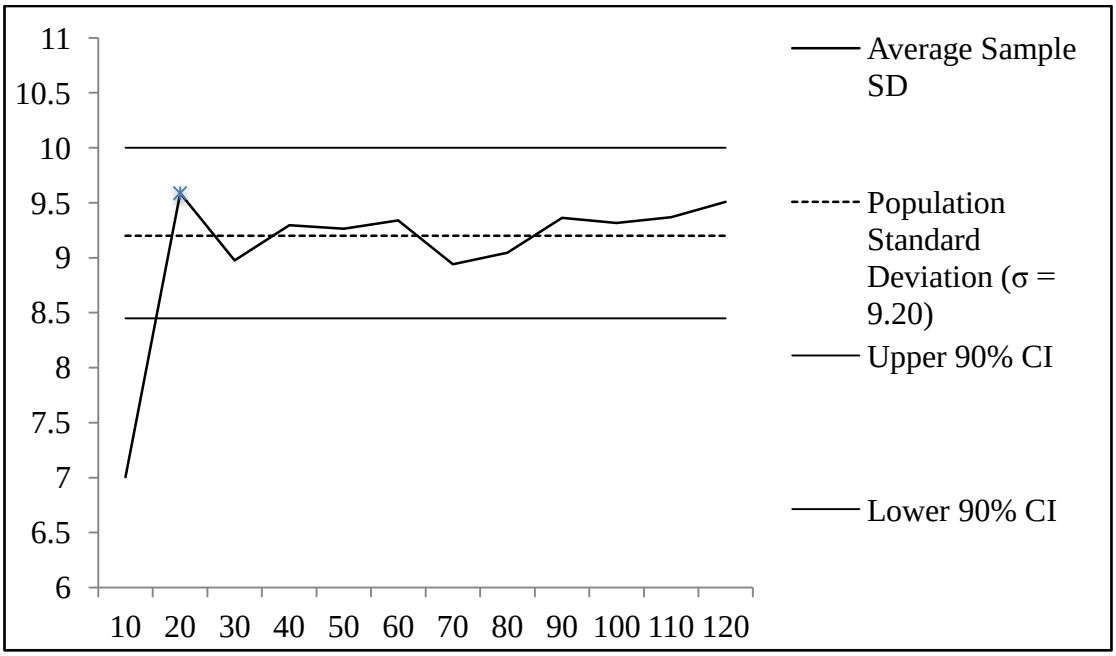

Figure 3.11. TMT A sample standard deviations with 90% confidence intervals

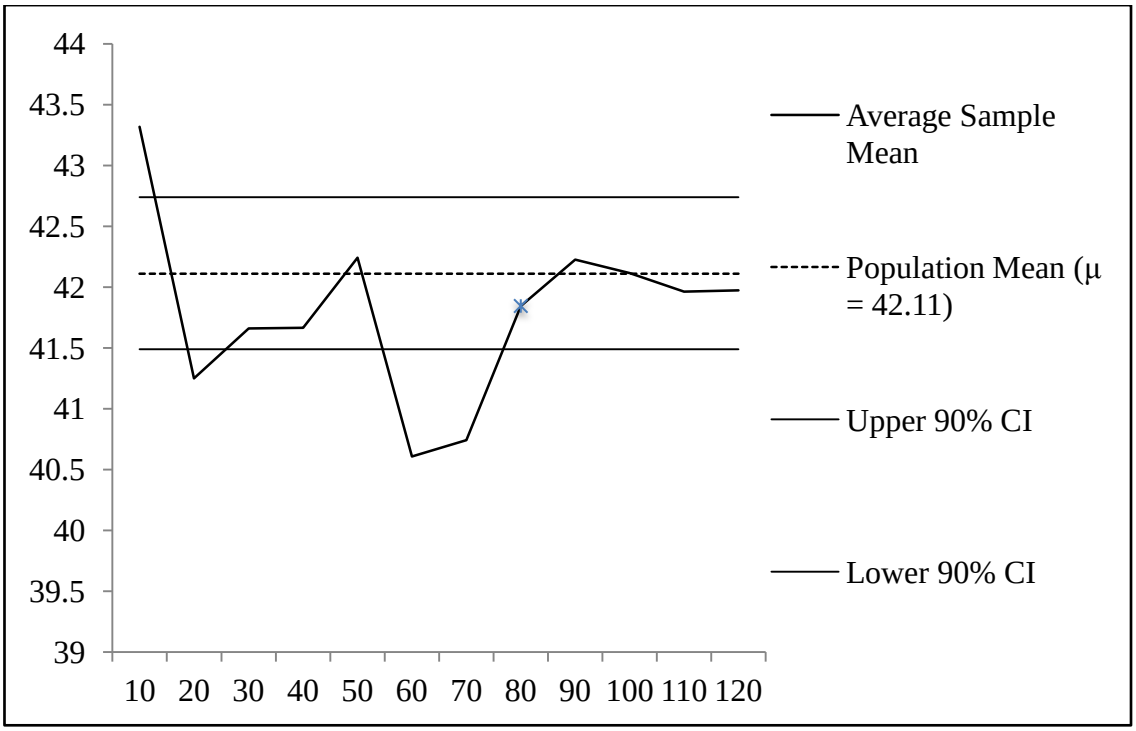

Figure 3.12. COWAT sample means with 90% confidence intervals

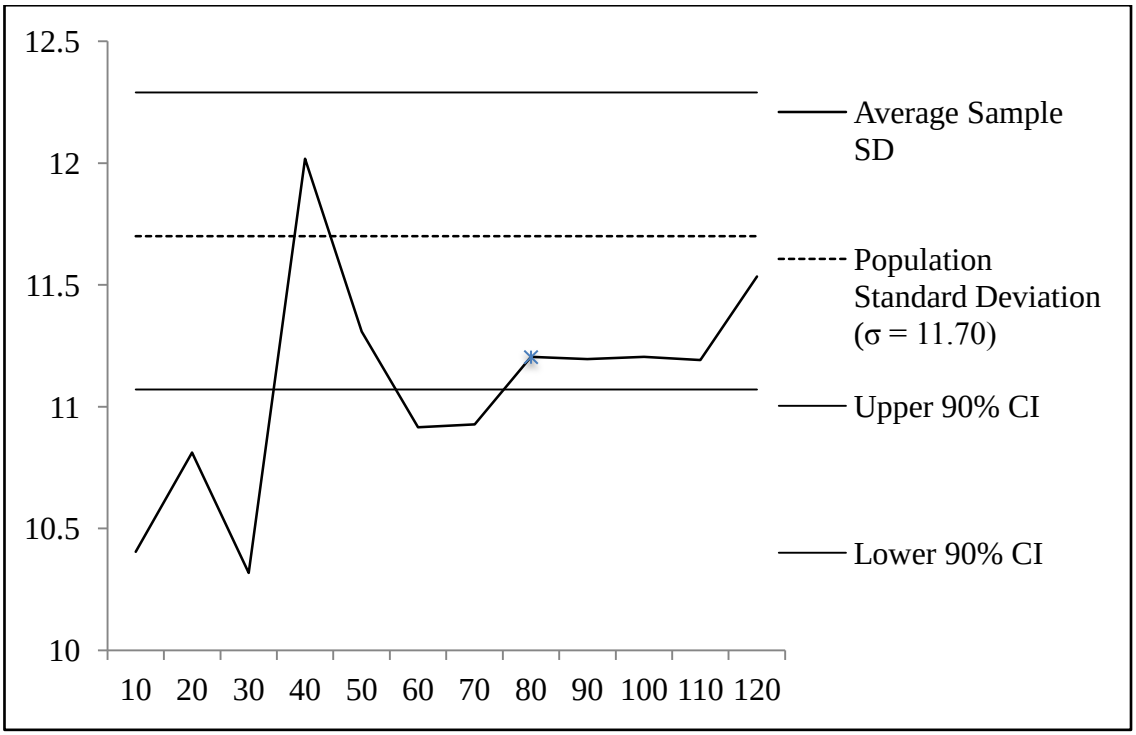

Figure 3.13. COWAT sample standard deviations with 90% confidence intervals

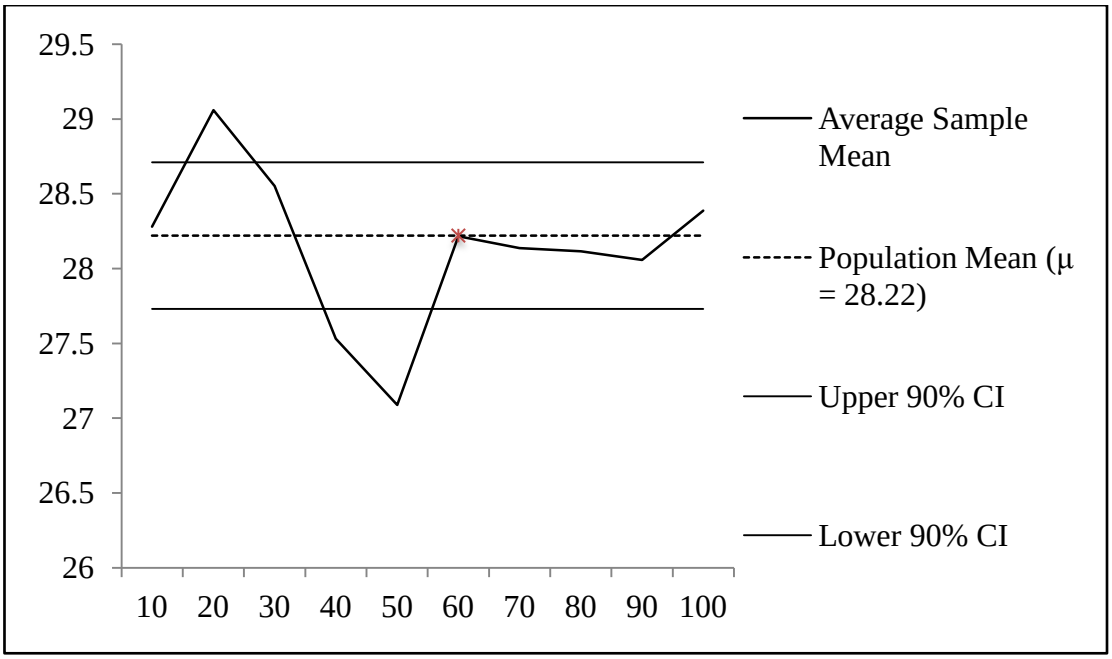

Figure 3.14. WAIS-III Symbol Search sample means with 90% confidence intervals

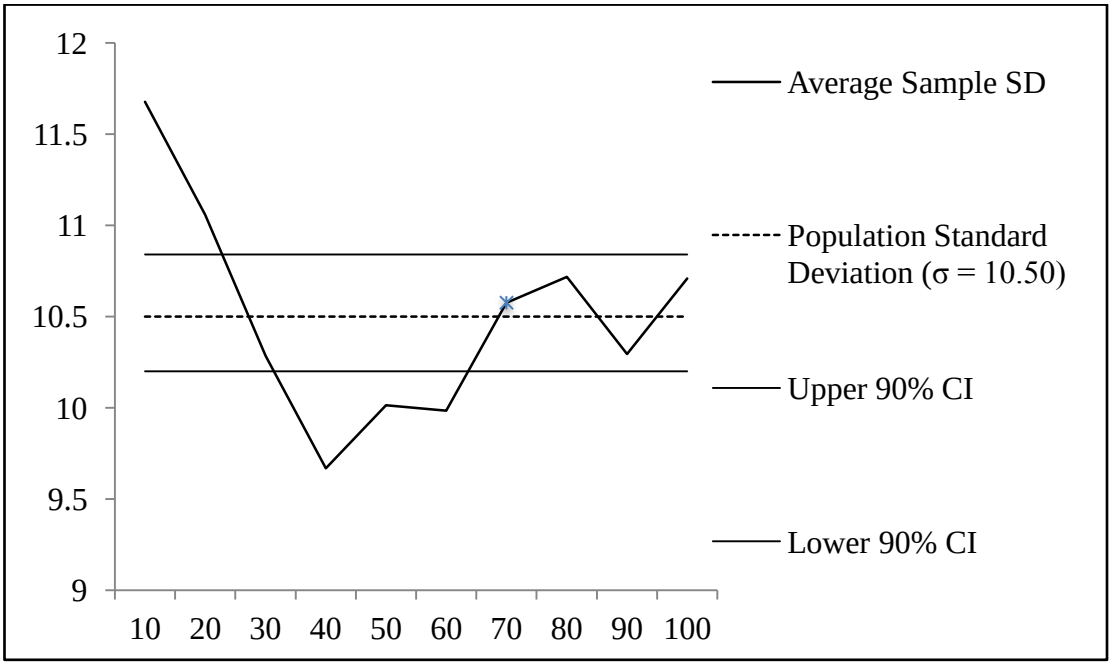

Figure 3.15. WAIS-III Symbol Search sample standard deviations with 90% confidence intervals

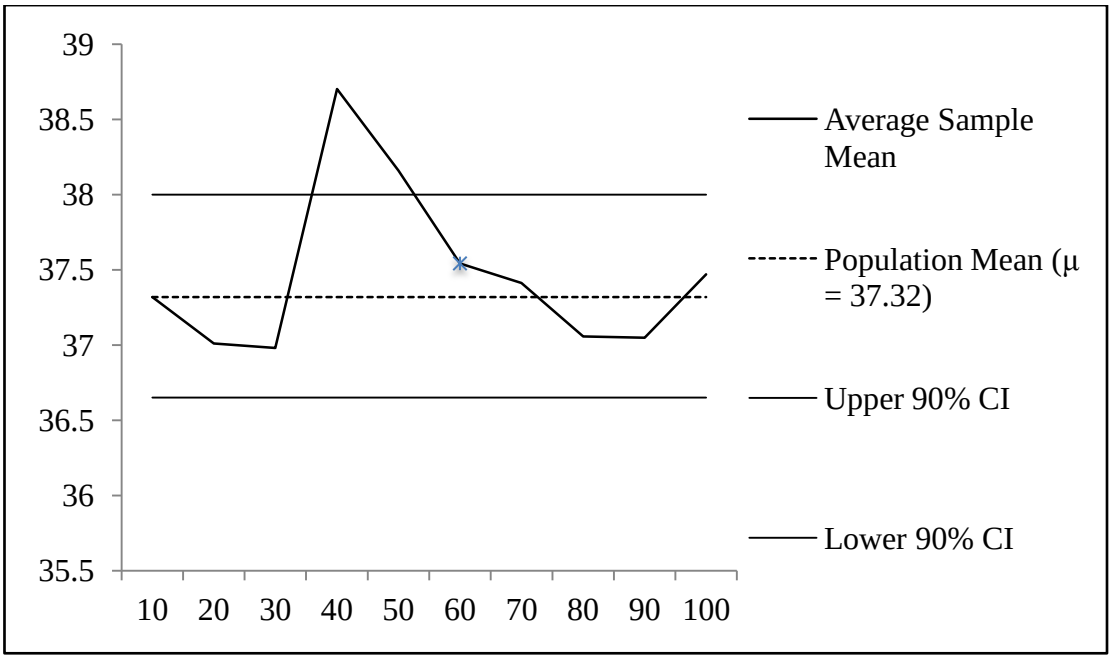

Figure 3.16. WTAR sample means with 90% confidence intervals

Figure 3.17. WTAR sample standard deviations with 90% confidence intervals

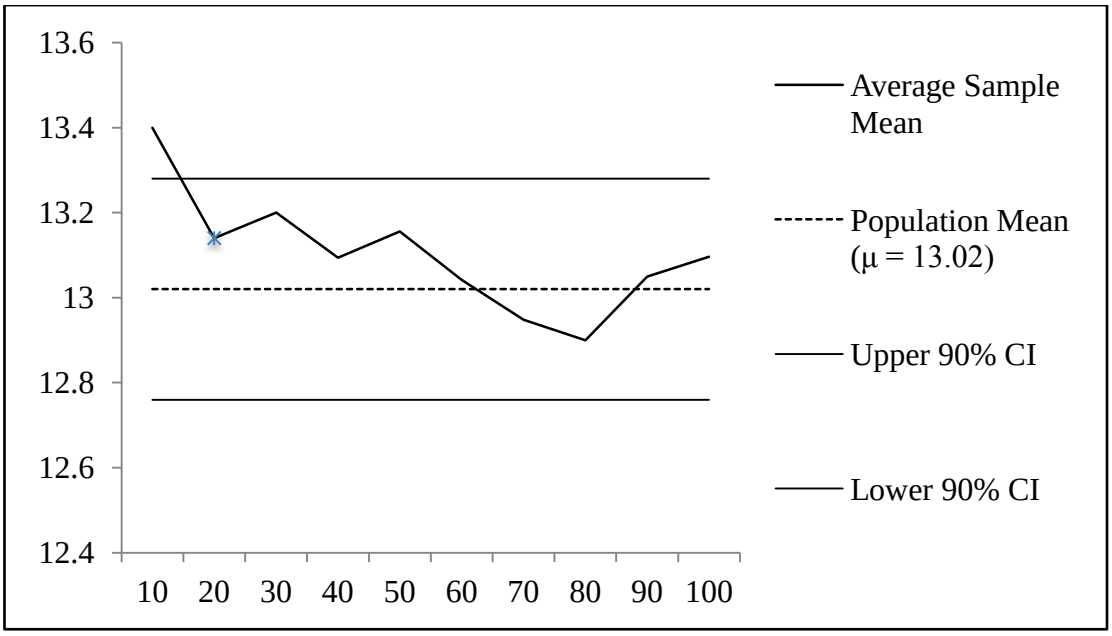

Figure 3.18. Rey 15 Item sample means with 90% confidence intervals

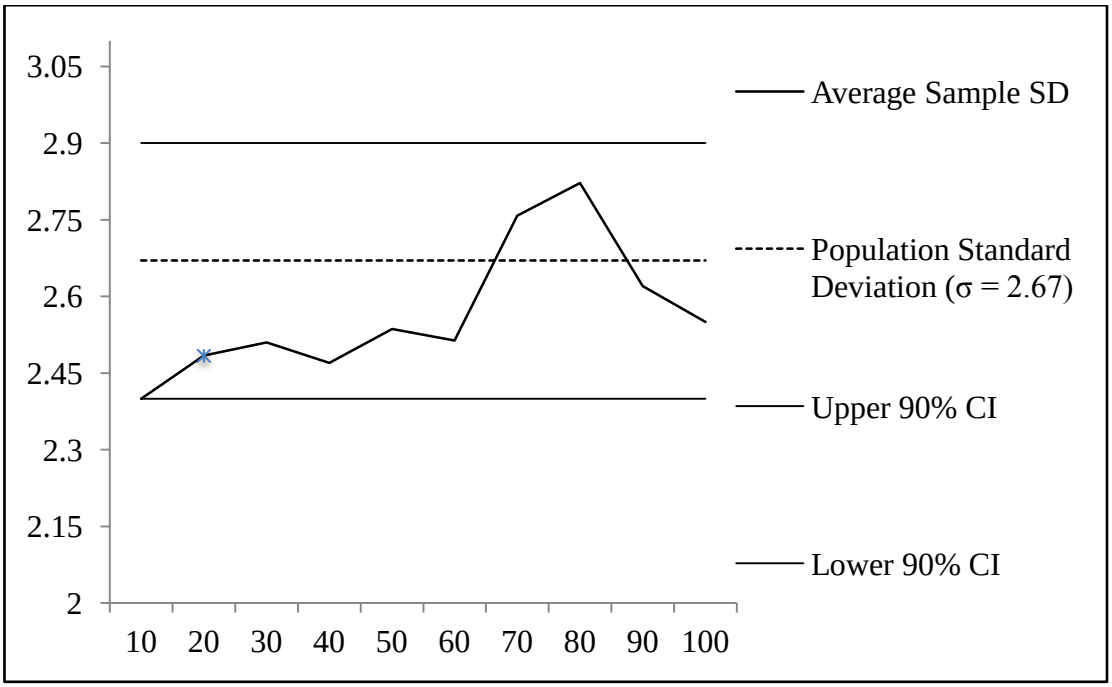

Figure 3.19. Rey 15 Item sample standard deviations with 90% confidence intervals

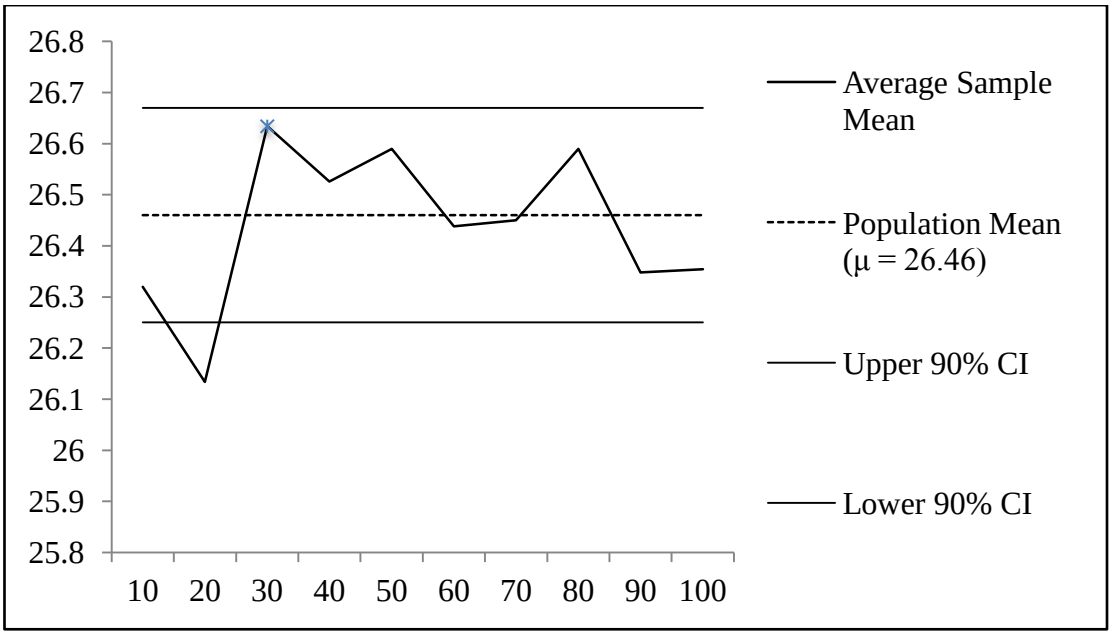

Figure 3.20. HVOT sample means with 90% confidence intervals

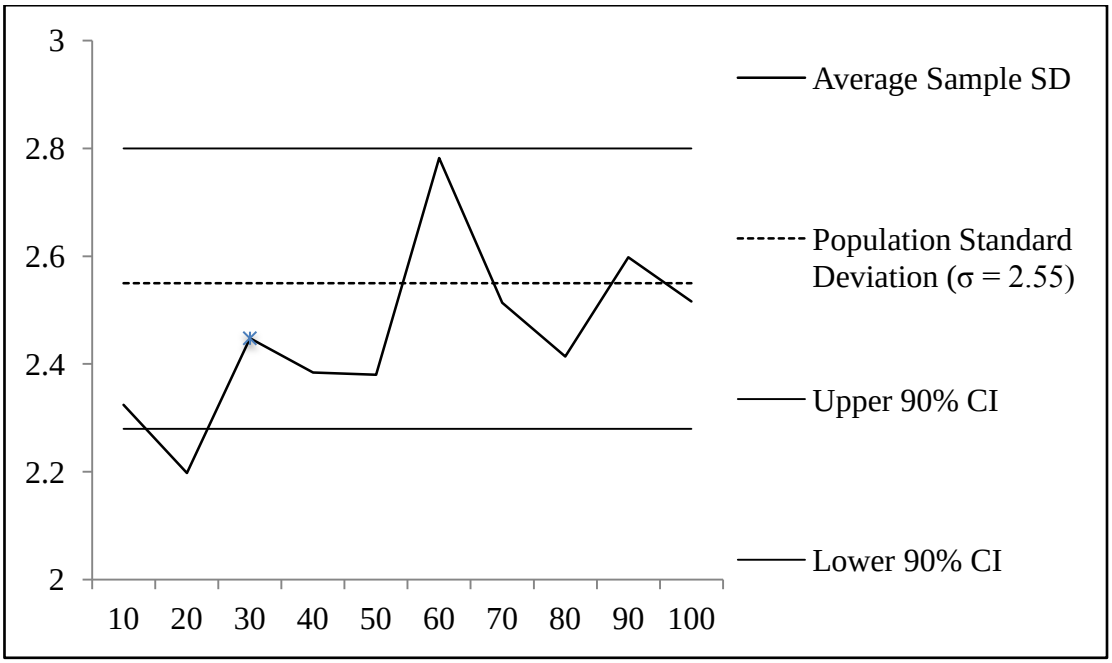

Figure 3.21. HVOT sample standard deviations with 90% confidence intervals

All figures above were examined in order to determine at what sample size the sample means and standard deviations stabilise. Table 3.12 summarises this information. The minimum sample size needed for each test is determined after considering both the mean and standard deviation together.

Table 3.12.
Sample Sizes needed for Stable Means and Standard Deviations for Seven Tests

	Sample Size Required		
Test	*M*	*SD*	Minimum
TMT B	60	50	60
TMT A	80	20	80
COWAT	80	80	80
WAIS-III SS	60	70	70
WTAR	60	40	60
Rey 15 Item	40	20	40
HVOT	30	30	30

The data from Table 3.12 were then plotted with regard to degree of skewness. A trendline or line of best fit was added between the series data points, as displayed in Figure 3.22. A fourth order polynomial was selected as the best fit for the data ($R^2 = 0.995$).

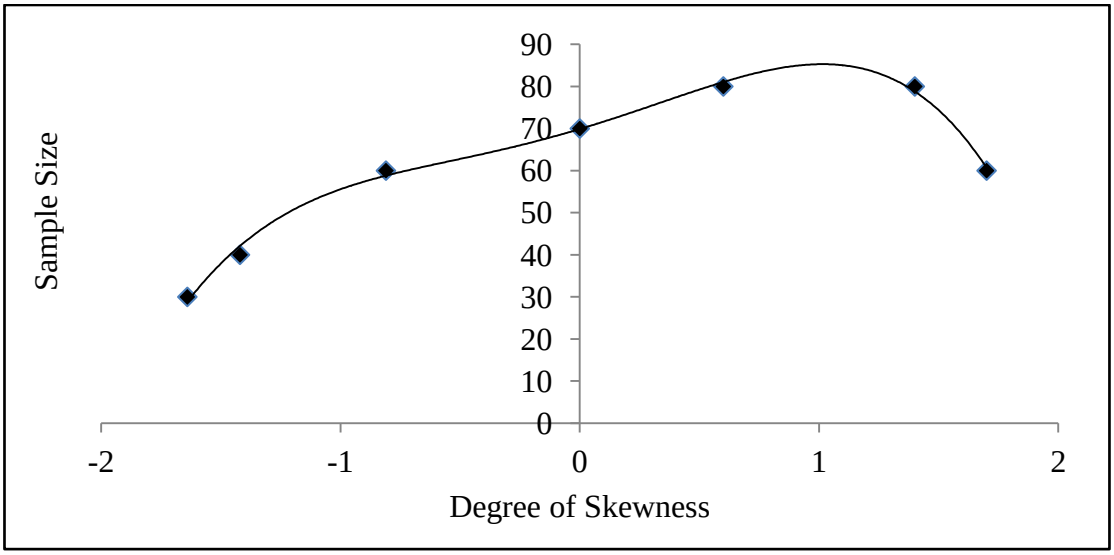

Figure 3.22. Stable means and standard deviations as a function of skewness

The trendline formula to calculate optimum sample size needed for tests with varying degrees of skewness was:

$$y = -5.3239\,x^4 - 2.1725\,x^3 + 5.8663\,x^2 + 17.028\,x + 69.869 \quad \text{Formula 11}$$

Where:

x = the skewness statistic of the sample distribution

This formula can then be used to calculate the optimum sample size needed for normative data for psychometric tests. For cross-validation purposes and to

demonstrate the effectiveness of Formula 11 examinations of two other measures from the WAIS-III Standardisation sample that reflected different degrees of skewness were conducted.

3.3.1 Cross Validation

Following the same methodology, the demographic characteristics and skewness statistics, were calculated for two further tests from the WAIS-III standardisation study. The tests chosen were the WAIS-III Information and the WAIS-III Digit Symbol – Symbol Copy (DSC) subtests. Both are subtests from the WAIS-III (Wechsler, 1997a) with Information designed to access the test-taker's general knowledge of literature, geography, science, and history and Digit Symbol – Symbol Copy assessing the test-taker's ability to copy a abstract symbol. Table 3.13 presents demographic characteristics for the two tests while Table 3.14 presents the normality test statistics. .

Table 3.13.
Descriptive Statistics for WAIS-III Information and Digit Symbol – Symbol Copy

							Gender	
Test	*N*	*M*	*SD*	*Mdn*	Age *M* (*SD*)	Education *M* (*SD*)	M	F
Information	1250	15.44	5.53	16	48.36 (23.96)	12.47 (2.63)	581	669
Digit Symbol – Copy	1250	103.38	27.6	108	48.36 (23.96)	12.47 (2.63)	581	669

Table 3.14.
Tests of Normality for WAIS-III Information and Digit Symbol – Symbol Copy

					Statistics	
Test	Skewness	*SE*	Bulmer	*SE*	Statistic	Sig.
Information	0.05	0.07	Normal	0.14	0.08	0.00
Digit Symbol - Copy	-0.74	0.07	Moderately	0.14	0.14	0.00

Note: Statistics = Kolmogorov-Smirnov Statistics; Bulmer refers to Bulmer's 1979 Classification; Test 1 = WAIS-III Information; Test 2 = WAIS-III Digit Symbol – Symbol Copy.

Histograms displaying underlying distributions of the two tests were also produced and are displayed in Figures 3.23 (Information) and 3.24 (Digit Symbol-Copy).

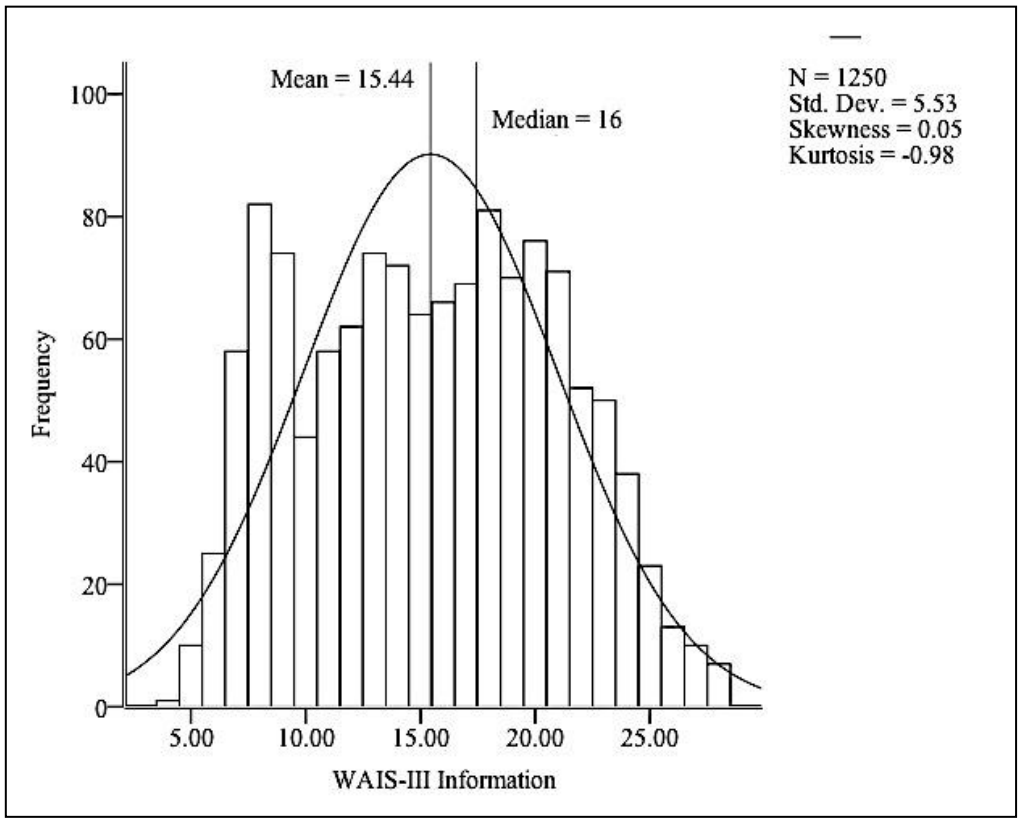

Figure 3.23. Histogram of WAIS-III Information distribution

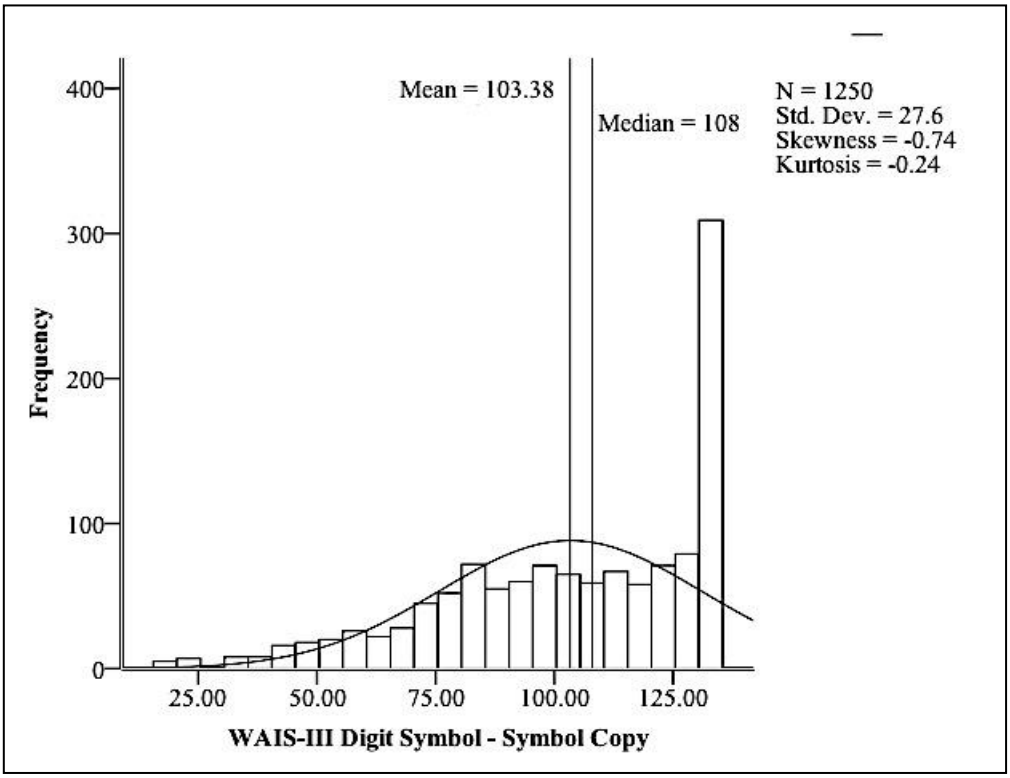

Figure 3.24. Histogram of WAIS-III Digit Symbol –Copy distribution

Formula 11 was then applied to each test substituting the skewness statistic into the equation. For the Information subtest:

$y = -5.3239\ (0.05)^4 - 2.1725\ (0.05)^3 + 5.8663\ (0.05)^2 + 17.028\ (0.05) + 69.869$
$= 70.73$

For Digit Symbol – Copy:

$y = -5.3239\ (-0.74)^4 - 2.1725\ (-0.74)^3 + 5.8663\ (-0.74)^2 + 17.028\ (-0.74) + 69.869$
$y = 59.76$

As can be appreciated from the above calculations, WAIS-III Information with a 0.05 degree of skewness (i.e., normally skewed) will need a sample size of approximately 70 in order to produce a stable mean and standard deviation. Whereas, the moderately negatively skewed distribution of the WAIS-III Digit Symbol – Symbol Copy would need a sample size of approximately 60 for a stable mean and standard deviation. In order to cross-validate these approximations, SPSS 21 was used to randomly draw cases with replacement from their respective distributions with sample sizes of 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100. Samples were drawn five times each and the averages calculated (see Table 3.15). Ninety percent confidence intervals were also calculated for the mean and standard deviations using Formula 9 and 10 and Bootstrapping, respectively and these are presented in Table 3.16.

Table 3.15.
Descriptive Statistics for WAIS-III Information and Digit Symbol – Symbol Copy as a Function of Sample Size

Test	*n*	μ	σ	Sample *M*	Sample *SD*
WAIS- III Information	10	15.44	5.53	17.20	5.59
	20	15.44	5.53	14.64	5.12
	30	15.44	5.53	15.53	5.42
	40	15.44	5.53	15.97	5.63
	50	15.44	5.53	14.94	5.78
	60	15.44	5.53	15.41	5.39
	70	15.44	5.53	15.22	5.57
	80	15.44	5.53	15.28	5.52
	90	15.44	5.53	15.27	5.47
	100	15.44	5.53	15.32	5.60
WAIS-III Digit Symbol -	10	103.38	27.6	105.20	28.64
Symbol Copy	20	103.38	27.6	102.37	27.70
	30	103.38	27.6	102.54	29.44
	40	103.38	27.6	103.57	26.58
	50	103.38	27.6	101.57	28.48
	60	103.38	27.6	104.48	27.02
	70	103.38	27.6	103.87	27.00
	80	103.38	27.6	103.22	27.86
	90	103.38	27.6	103.50	28.23
	100	103.38	27.6	103.52	27.36

Table 3.16.
90% Confidence Intervals for Population Means and Standard Deviations

		90% Confidence Intervals			
		M		*SD*	
Test	*SEM*	Upper	Lower	Upper	Lower
1	0.79	15.70	15.18	5.66	5.39
2	0.16	104.7	102.08	28.52	26.72

Note: Test 1 = WAIS-III Information; Test 2 = WAIS-III Digit Symbol – Symbol Copy

Using the information from Tables 3.15 and 3.16, graphs were generated in order to determine the sample size at which the mean and standard deviation stabilises for each test. These are presented in Figures 3.25 through 3.28.

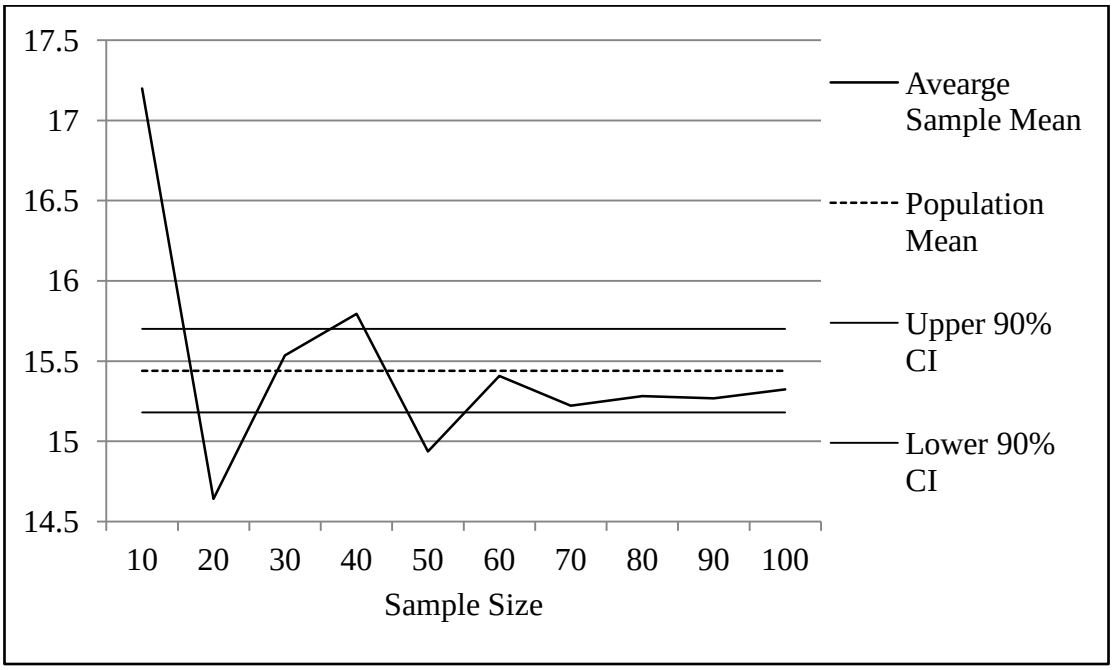

Figure 3.25. WAIS-III Information sample means with 90% Confidence Intervals

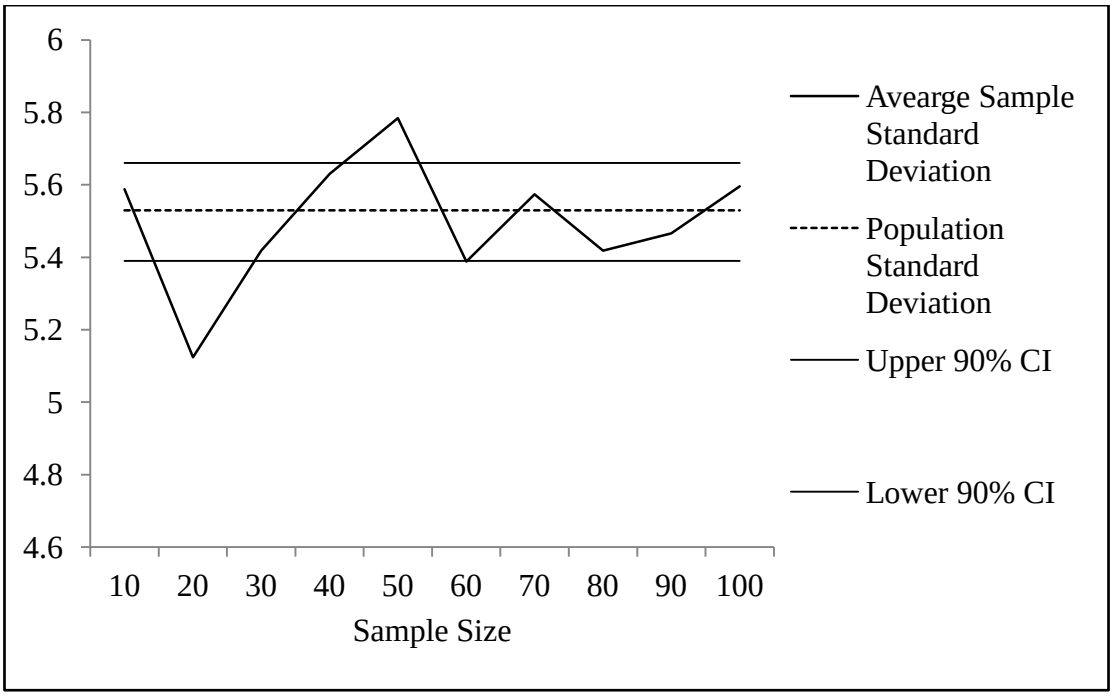

Figure 3.26. WAIS-III Information sample standard deviations with 90% Confidence Intervals

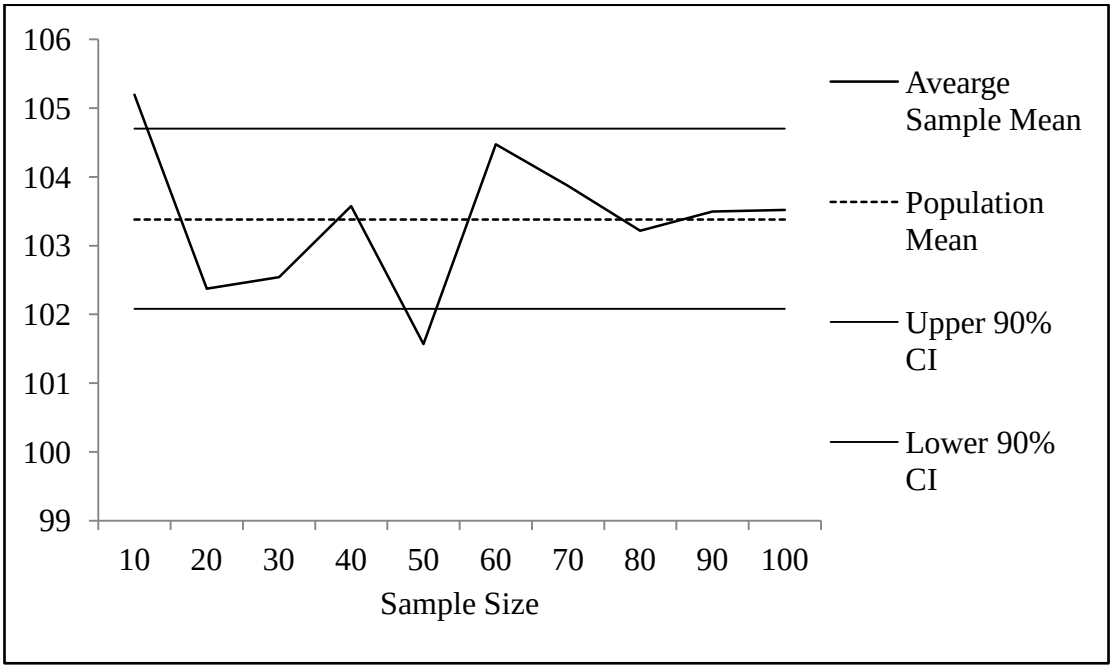

Figure 3.27. WAIS-III Digit Symbol – Symbol Copy sample means with 90% Confidence Intervals

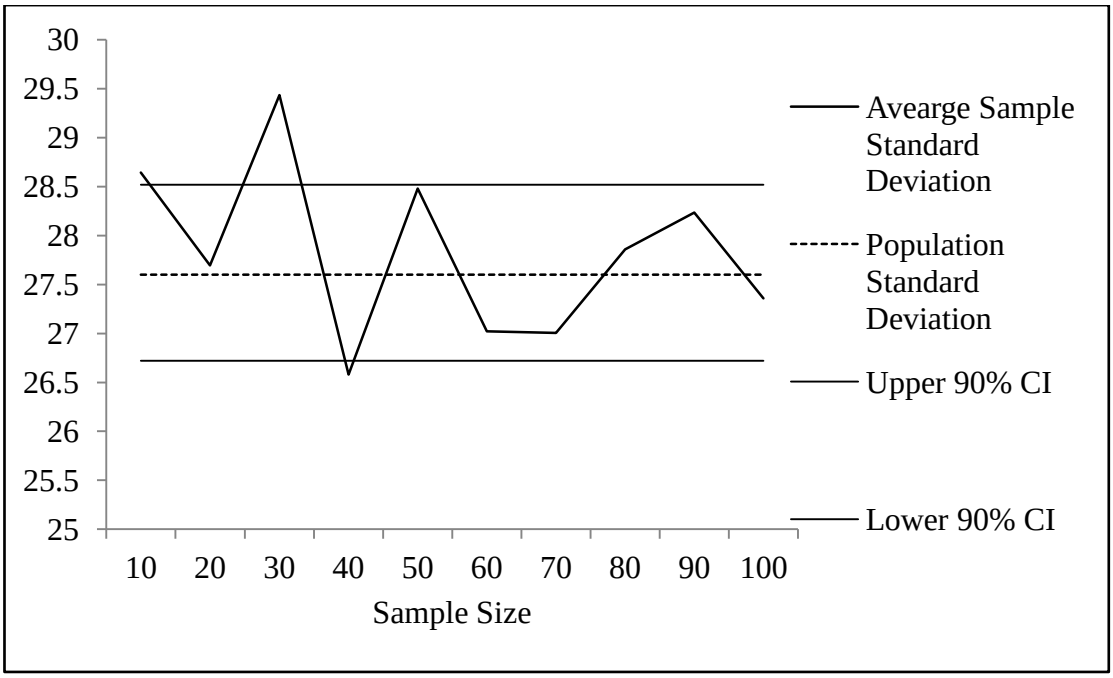

Figure 3.28. WAIS-III Digit Symbol – Symbol Copy sample standard deviations with 90% Confidence Intervals

As can be seen in the above figures, WAIS-III Information has a stable mean and standard deviation at a sample size of 70. WAIS-III Digit Symbol – Symbol Copy has a stable mean and standard deviation at sample size 60. These are consistent with the estimated sample sizes calculated using Formula 11 above and suggests that the formula 11 is an adequate method to predict required sample sizes for distributions with different degrees of skewness.

3.4 Summary

Normative samples are the foundations of clinical interpretation. They represent the theoretical population through which comparison with a client's performance can be judged. Normative samples are collected through a variety of methods. Census-based norms are collected to represent the demographic variables of a given country's census. The disadvantage is that it may under-represent minority groups, such as specific ethnic groups or extremes of demographic characteristics that while uncommon in the general population present specific confounds with respect to clinical inferences. This has the potential of introducing error in the interpretation process. The recruitment method is applied when a researcher uses a set of criteria to select participants to make up the normative sample. However, this method will not be representative of the census population parameters. Anchor-norms are the most commonly employed in neuropsychology as a target audience is selected and tested in order to meet the needs of the clinician or the researcher. This method is particularly useful for testing minority groups or populations with specific impairments. The disadvantage of collecting norms on such specific target populations is that sample sizes are often small.

Error is introduced into the interpretation process when using normative samples. Clinicians must consider whether the normative sample they are using as a comparison group for their client is representative particularly with reference to salient demographic variables such as age, gender, level of education, or ethnicity. It would be an unethical and thoughtless clinician who compared a 60-year-old

client's scores with a normative sample of 18-year-olds and concluded the presence of a cognitive impairment.

Another important issue when evaluating the appropriateness of normative samples is sample size. While it is commonly reported that normative samples should have a minimum of 50 cases per cell, little research has been conducted to establish any recommended minimum and in particular, the work of Bridges and Holler (2007) has highlighted the importance of considering skewed distributions. This is particularly important in neuropsychology where small sample sizes and skewed distributions abound. While methods such as meta-norming were developed to overcome some of the issues associated with small sample sizes in normative data, the problem remains. Accordingly, the purpose of Study Two was to systematically evaluate optimal sample size with regard to creating stable means and standard deviations while considering the impact of skewness.

The study consisted of using established normative data on seven commonly used neuropsychological tests with sample sizes greater than 200. Due to the large sample sizes, the means and standard deviations for each test were treated as population parameters. Sample metrics were then calculated for sample sizes ranging form 10 to 120 cases and compared to the population parameter. The sample size, at which the sample parameters converged to within the 90 percent confidence interval of the population parameters, indicated stable measures of central tendency and variance.

The current study found that the optimal *n* is not 50, and that the sample sizes required for stable measures were influenced by skewness. Interestingly, for normally distributed data, the sample size required for stable measures of central tendency and variance was 70. For negatively skewed distributions, the sample size ranges from $n = 30$ to $n = 70$ and for positively skewed distributions the required sample sizes were between 60 and 85. In the absence of computations of skewness in the normative data for the tests they employ, it is highly recommended that clinicians use normative samples with at least 90 cases in each cell. If skewness statistics are available Formula 11 can be used to calculate the optimal sample size required for stable measures of central tendency and variance. It should be noted that this formula is no means absolute in its calculation of sample size as there is a potential for it to be dependent on the samples and tests used to generate it. Regardless, it does serve as a valuable resource for clinicians and researchers alike and should be extended upon by inclusion of additional neuropsychological measures and normative distributions.

Overall, Chapter Three has evaluated the psychometric issues associated with using normative studies. This is the first stage in the clinical interpretation process. The next level involves understanding and determining whether a score falls within the impaired range and is indicative of abnormality and this is the focus of the next chapter.

CHAPTER FOUR
ASSESSING ABNORMALITY

4.1 Levels of Abnormality

It is important for clinicians to understand how one determines whether a standardised score reflects abnormality or impairment and can be considered in two ways. The first is the degree of abnormality for the individual score compared to the normative sample. This point estimate compares an observed score with that of the standardised distribution of the test. The second is the intra-individual comparisons of two tests. At this level, inferential statistics are employed to determine whether the standardised scores on the two tests differ significantly. It is important to note here a common clinical misconception: a significant difference between two scores does not signal abnormality, it merely indicates that the two numbers are not the same. While significance is necessary for abnormality, it is not sufficient. Abnormality fundamentally relates to the frequency or rarity of the magnitude of the difference between two scores, with cut-offs of 10 or five percent most commonly encountered in clinical practice.

4.2 Abnormality at the Individual Test Level

A linear standardised score allows a clinician to compare an individual with a normative sample in order to make inferences about performance. As described earlier, the most common method employed by clinicians is to convert the raw score into a *z* score using the mean and standard deviation of the sample and then consult the area under the curve of the normal distribution to indicate the percentile rank. Clinicians are then able to compare this score with a predetermined cut-off score and ascertain whether the individual's score is lower than would be expected in the normative sample. Base rates are preset at the discretion of the clinician but are conventionally placed at either the 5% or 10% level, with corresponding *z* scores of approximately -1.6 and -1.27 respectively. Many researchers have opposed the use of *z* scores because the methodology treats the statistics as parameters of the population rather than sample statistics and consequently increases the chance of Type I errors (Crawford & Howell, 1998; Crawford & Garthwaite, 2002; Crawford, Garthwaite, Howell, & Gray, 2004; Crawford, Garthwaite, Azzalini, Howell, & Laws, 2006; Crawford & Garthwaite, 2012). Type I errors result in the clinician incorrectly classifying individuals as impaired when in reality they are not. However, this standardisation method fails to measure the degree of abnormality found in the test. As Crawford, Garthwaite, and Gault (2007) note, "...information on the rarity or abnormality of test scores (or test score differences) is fundamental in interpreting the results of a neuropsychological assessment" (p. 419). Some methods have been proposed as alternatives to the *z* score and these will be evaluated in terms of their vulnerability to sample size and sensitivity to Type I errors.

4.2.1 Solutions for Determining Abnormality at the Individual Test Level

Crawford and Garthwaite (2012) evaluated six different methods for comparing an individual to a control. Five methods employed *t*-distributions while the sixth approach used the conventional *z* score. By running Monte Carlo simulation trials, three of the inferential methods were deemed obsolete due to substantially high error rates especially with an increase in sample size. Z scores were noted to increase in Type I errors as the sample size decreased. The two remaining statistical methods were Crawford and Howell's (1998) "*t*-test approach"

and the prediction interval method (Barton, Press, Keenan, & O'Connor, 2002). The *t*-test approach (Crawford & Howell, 1998) is computed using Formula 12 below.

$$t = \frac{X_1 - \overline{X_2}}{s_2\sqrt{\frac{N_2+1}{N_2}}} \qquad \text{Formula 12}$$

Where:

X_1 = *the observed score*
$\overline{X_2}$ = *the normative sample mean*
s_2 = *the standard deviation of normative sample*
N_2 = *sample size*

This method estimates the abnormality of an obtained score and compares it for significance against the scores of the control sample. The *p* value is used to test this significance, and also acts as an indicator of the proportion of the control population who would obtain a lower score (Crawford & Howell, 1998).

The prediction interval method (Formula 13) is used to calculate the standard error between the sample mean and an additional case (Crawford & Garthwaite, 2012). This is then multiplied by *t* corresponding to *n* – 1 degrees of freedom providing a prediction interval on the control mean.

$$95\%\ \text{PI} = \bar{x} \pm t_{n-1,\ 0.975}\left(s\sqrt{\frac{n+1}{n}}\right) \qquad \text{Formula 13}$$

Where:

m = the number of people scoring below the given score
k = the number of people obtaining the given score

Crawford and Garthwaite (2012) explained how these methods while quite similar, have some distinct differences. When testing for abnormality, the *t*-test approach generally uses a one-tail test of significance while the prediction interval method generally uses a two-tailed test of significance. Additionally, the *t*-test approach provides a probability and has supplementary statistics developed by Crawford and Garthwaite (2002), which provide 95% confidence intervals around the point-estimate. Statistically, however, these two methods produce identical results when Monte Carlo simulation trials are run. For example, for a sample size of 10, both methods have Type I error percentages of 5.01 (Crawford & Garthwaite, 2012).

The benefit of using the *t*-test approach, in particular, is that it can be applied to small sample sizes as an alternative to the *z* score. For example, Crawford and Howell (1998) suggest "that the modified *t*-test be used with an *N* of less than 50" (p. 485) and have even recommended that it "should be used in preference to *z*" (Crawford et al., 2006, p. 673). However, this method assumes normality of the underlying actual distribution, with concerns remaining that Type I errors become inflated when these distributions are skewed.

Crawford et al. (2006) aimed to test the effect of normal, skewed, and leptokurtic distributions on Type I error rates for the *z* score method and Crawford and Howell's (1998) approach. Using Monte Carlo simulation trials with an error rate preset at 0.05, results indicated that for small to moderate sample sizes (n = 5 to 20), Type I errors were larger for the *z* score method than Crawford and Howell's *t*-test approach. With extreme skewness and kurtosis in the distributions, the percentage of Type I errors was between 7.84% and 9.96% for the *t*-test approach and between 8.68% and 14.31% when using *z* scores. When there was no kurtosis

but extreme skewness in the distribution of small to moderate sample sizes, the Type I error percentage for the *t*-test approach was between 7.70% and 8.27% and between 8.85% and 13.37% for *z* scores. For small to moderate sample sizes, however, with distributions that have a moderate skew and no leptokurtosis, the percentage of Type I errors was 5.88% to 5.93% and 7.10% to 11.28% for the two methods, respectively.

These findings highlight the use of the Crawford and Howell (1998) method over the *z* score method with normal, skewed, or leptokurtic distributions of small to moderate sample sizes. Similar to previous research, results of this study also found the Crawford and Howell's method had a lower percentage of Type I errors than *z* scores across different sample sizes ($n = 5$ to $n = 100$). Furthermore, although Type I errors were inflated when using Crawford and Howell's method, the researchers did not believe these were drastic and overall, they still recommended this method for use by clinicians.

4.3 Abnormality Between Two Tests and Solutions

The second consideration of clinicians in neuropsychological assessments is to analyse the difference between scores on two tests and calculate whether this discrepancy is abnormal. The methods developed by Payne and Jones (1957) and Crawford, Howell, and Garthwaite (1998) both provide viable approaches to this task. The Payne and Jones method determines the abnormality of the difference between two scores and the magnitude of the discrepancy compared to the population by the following formulae (Formula 14):

$$Z = \frac{x_1 - x_2}{\sqrt{S1^2 S2^2 - 2r_{xy}}} \qquad \text{Formula 14}$$

When standardised scores are used this simplifies to

$$Z = \frac{Z_x - Z_y}{\sqrt{2 - 2r_{xy}}} \qquad \text{Formula 15}$$

Where:

r_{xy} = the correlation between the two tests
Z_x = the standardised score for test one
Z_y = the standardised score for test two
x_1 = score for test one
x_2 = score for test two
s_1 = standard deviation for test one
s_2 = standard deviation for test two

This simplified version clearly indicates that the determination of the frequency of a difference between two scores is a product of the magnitude of the difference and the correlation between them.

Although the practicality of this method has been demonstrated in clinical neuropsychology (Ley, 1972), it does however treat the sample statistics as if they are population parameters (Crawford & Garthwaite, 2002). Furthermore, clinicians must know the correlations between the two tests in question and this is not always readily available to clinicians. The alternative method developed by Crawford, Howell, and Garthwaite (1998) calculates abnormality between two tests based on sample statistics. This modified *t*-test (formula 16) is also a useful method when used with small sample sizes.

$$t = \frac{Z_x - Z_y}{\sqrt{(2 - 2r_{xy})((N_2 + 1)/N_2)}} \qquad \text{Formula 16}$$

Where:

Z_x = *the standard score for test one*
Z_y = *the standard score for test two*
r_{xy} = *the correlation between the two tests*
N_2 = *the number of persons in the sample*

Multiplying the *t*-score from this formula by 100 provides the point-estimate of the abnormality of the difference between the scores. Crawford and Garthwaite (2002) also provide confidence limits for this point-estimate method. As with the Payne and Jones approach, the Crawford et al. (1998) method also requires the clinician to have access to the correlation between the two tests in question.

4.4 Summary

Chapter four addressed the second stage of the clinical decision making process. After standardising a raw score, a clinician must determine whether a standardised score is indicative of abnormality. This can occur on two levels: at the basic individual test level, and in relation to comparisons between two tests. The chapter summarised both of these levels and discussed potential errors and current strategies available to correct them. While these methods are practical for clinicians to employ in clinical decision-making, they fail to address or account for the influence of skewed distributions. That is, the standardised scores and the methods for determining rates of abnormality at the individual test score level all assume normality of the actual distribution. Therefore the next chapter will empirically evaluate the interpretation and issues involved when standardising a raw score with a skewed distribution.

CHAPTER FIVE
SKEWED DISTRIBUTIONS AND CLINICAL DECISION MAKING

5.1 Interpretation Issues When Standardising on Skewed Distributions

As demonstrated in Chapter three, the benefit of using normalised scores is that it allows clinicians to correctly take into account the skewness of the underlying normative distribution in representing percentiles. Donnell et al. (2011) took this a step further by studying the actual interpretative effect of using linear versus normalised scores on negatively skewed neuropsychological test data. This study used archival data from the Vietnam Experience Study of 4462 randomly selected US Army veterans, who had served during the Vietnam War era. All participants undertook three days of evaluation including completing a comprehensive neuropsychological battery (i.e., nine neuropsychological tests, from which 21 variables were derived). An analysis of the normative samples' distribution identified eight variables that had skewed distributions. These variables were transformed into a normal distribution. The remaining normally distributed variables were converted to linearly transformed *z* scores and then assigned to corresponding scaled scores.

The researchers then measured the degree of difference between the normalised and linear scaled scores of one of the skewed neuropsychology tests and found that some linear scaled scores were within the abnormal range while the corresponding normalised scaled score fell in the normal range, thus altering the interpretation depending upon the approach that was used (Donnell et al., 2011). This was particularly evident at the lower end of the negatively skewed distribution. Donnell et al (2011) emphasised:

> "…the importance of knowing the performance frequency distributions of various tests before assuming the data are normally distributed and using strict linearly transformed standard scores. Without normalisation of the performance distributions, lower scores can easily be misinterpreted as being pathological when they may not be." (p. 1105)

Although this study is invaluable in demonstrating the potentially catastrophic impact of treating skewed distributions as if they were normally distributed, it does not address the degree of skewness. It is, therefore, necessary to determine how different skewed distributions affect the way data is interpreted. There is little value in belabouring clinicians with the perils of not knowing the underlying distribution of the tests they employ, if no guidelines exist for them to consider, accommodate, or adjust their practices for differing levels of skewness. Clinicians need to make educated decisions about how to approach standardisation in clinical practice. This concern forms the basis of the following study.

5.2 Study Three –Errors With Standardising Skewed Distributions

This study assessed the magnitude of errors produced when different standardisation methods were applied to raw score distributions with varied levels of skewness. It utilised the seven neuropsychological tests from Study Two with their broad range of skewed distributions. Three standardisation methods were analysed. These were the traditional linear *z* score transformation, Crawford and Howell's (1998) modified *t*-test and a new method created for this study, the Median *z* score transformation. This method is a modified version of the linear *z* score but is based on the median rather than the mean. It was hypothesised that using the median as a measure of central tendency instead of the mean in the *z* score transformation would

provide a standardised score more appropriate for highly skewed distributions as the median reflects the 50th percentile regardless of the degree of skewness. The formula for the Median *z* score transformation was as follows:

$$Mdn\ z = [x - mdn]/\ mdnsd \qquad \text{Formula 17}$$

Where:

x = *the observed score*
mdn = *the median*
mdnsd = *the median standard deviation*

The median standard deviation was a modified version of the standard deviation formula but reflected the spread of scores around the median.

$$Mdnsd = \sqrt{\frac{\sum(x-mdn)^2}{n-1}} \qquad \text{Formula 18}$$

Where:

x = *the observed score*
mdn = *the median*
n = *the population size*

It should be noted that a difference between the median and the mean is that deviations from the median do not sum to zero in a skewed distribution. The median standard deviations were calculated for each of the seven neuropsychological tests with the results presented in Table 5.1.

Table 5.1.
Medians and Median Standard Deviations for Seven Neuropsychological Tests

Tests	*N*	Median	Median Standard Deviation
COWAT	935	41	11.53
HVOT	379	27	2.6
TMT A	507	24	9.4
TMT B	507	55	23.14
Rey 15 Item Test	272	15	3.32
WAIS-III Symbol Search	1250	29	10.53
WTAR	389	39	8.22

Using the known raw scores for each percentile from the actual raw score distribution of each test, the three standardisation methods were then applied and compared for each distribution. For each method, the *z* score or *t*-score was calculated and was converted into the corresponding percentile rank using Table 2.4. This percentile rank was then compared to the original percentile rank, and the difference between them was calculated. Tables 5.2 through 5.8 present the data for each of the seven tests. It should be noted that the 5th and 10th percentiles are the most relevant for inferring impairment in neuropsychological assessment.

Table 5.2.
COWAT – Comparisons of Three Standardisation Methods

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-2.06	2	1	-1.99	2	1	-2.06	2	1
5	-1.46	7	2	-1.39	8	3	-1.46	8	3
10	-1.21	12	2	-1.13	13	3	-1.21	12	2
25	-0.69	24	-1	-0.61	27	2	-0.69	25	0
50	-0.09	46	-4	0	50	0	-0.09	47	-3
75	0.59	73	-2	0.69	76	1	0.59	72	-3
90	1.19	88	-2	1.3	91	1	1.19	88	-2
95	1.7	97	1	1.82	96	1	1.7	96	1
99	2.81	99.4	0.4	2.94	99.6	0.6	2.81	99.5	0.5

Table 5.3.
HVOT – Comparisons of Three Standardisation Methods

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-3.79	0.03	-0.97	-3.92	0.03	-0.97	-3.79	0.03	-0.97
5	-2.14	2	-3	-2.31	1	-4	-2.14	2	-3
10	-1.16	13	3	-1.35	9	-1	-1.16	13	3
25	-0.38	35	10	-0.58	28	3	-0.38	36	11
50	0.21	58	8	0	50	0	0.21	59	9
75	0.60	73	-2	0.38	65	-10	0.60	73	-2
90	1.00	84	-6	0.77	78	-12	1.00	84	-6
95	1.19	88	-7	0.96	84	11	1.19	88	-7
99	1.39	92	-7	1.15	88	-11	1.39	92	-7

Table 5.4.
TMT A – Comparisons of Three Standardisation Methods

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-1.51	6	5	-1.27	10	9	-1.51	7	6
5	-1.19	12	7	-0.96	17	12	-1.19	13	8
10	-1.08	14	4	-0.85	20	10	-1.08	14	4
25	-0.70	24	-1	-0.48	32	7	-0.70	24	-1
50	-0.21	42	-8	0	50	0	-0.21	42	-8
75	0.44	67	-8	0.64	74	-1	0.44	67	-8
90	1.32	91	1	1.5	94	4	1.32	91	1
95	1.92	97	2	2.08	98	3	1.92	97	2
99	3.67	99.98	0.98	3.80	99.98	0.98	3.68	99.98	0.98

Table 5.5.
TMT B – Comparisons of Three Standardisation Methods

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-1.52	6	5	-1.29	10	9	-1.52	6	5
5	-1.21	12	7	-0.99	16	11	-1.21	12	7
10	-0.99	16	6	-0.77	22	12	-0.99	16	6
25	-0.64	26	1	-0.43	33	8	-0.64	26	1
50	-0.20	42	-8	0	50	0	-0.20	42	-8
75	0.41	67	-8	0.61	73	-2	0.41	67	-8
90	1.20	88	-2	1.38	92	2	1.20	88	-2
95	2.11	98	3	2.27	99	4	2.12	98	3
99	3.69	99.98	0.98	3.82	99.98	0.98	3.69	99.98	0.98

Table 5.6.
Rey 15 Item – Comparisons of Three Standardisation Methods

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-3.10	0.02	-0.98	-3.09	0.2	-0.8	-3.11	0.1	-0.9
5	-2.39	1	-4	-2.52	1	-4	-2.39	1	-4
10	-1.51	6	-4	-1.81	4	-6	-1.51	7	-3
25	-0.38	35	-10	-0.90	19	-6	-0.38	36	11
50	0.74	77	27	0	50	0	0.74	78	28
75	0.74	77	2	0	50	-25	0.74	78	3
90	0.74	77	-13	0	50	-40	0.74	78	-12
95	0.74	77	-18	0	50	-45	0.74	78	-17
99	0.74	77	-22	0	50	-49	0.74	78	-21

Table 5.7.
WAIS-III Symbol Search – Comparisons of Three Standardisation Methods

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-2.21	1	0	-2.28	1	0	-2.21	1	0
5	-1.64	5	0	-1.71	4	-1	-1.64	5	0
10	-1.35	9	-1	-1.42	8	-2	-1.35	9	-1
25	-0.69	25	0	-0.76	22	-3	-0.69	25	0
50	0.07	53	3	0	50	0	-0.07	47	-3
75	0.65	74	-1	0.57	72	-3	0.65	75	0
90	1.31	91	1	1.23	89	-1	1.31	91	1
95	1.69	96	1	1.62	95	0	1.69	96	1
99	2.26	99	0	2.18	99	0	2.27	99	0

Table 5.8.
WTAR – Comparisons of Three Standardisation Methods

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-3.28	0.08	-0.92	-3.42	0.06	-0.94	-3.29	0.08	-0.92
5	-1.90	3	-2	-2.07	2	-3	-1.91	3	-2
10	-1.16	13	3	-1.34	9	-1	-1.16	13	3
25	-0.66	25	0	-0.85	20	-5	-0.66	26	1
50	0.21	42	-8	0	50	0	0.21	59	9
75	0.83	80	5	0.61	74	-1	0.83	80	5
90	1.20	88	-2	0.97	84	-6	1.20	88	-2
95	1.33	91	-4	1.09	87	-8	1.33	91	-4
99	1.58	95	-4	1.34	91	-8	1.58	95	-4

The difference scores for the 5th and 10th percentiles were then summarised for each of the seven tests, as presented in Table 5.9.

Table 5.9.
Summary of the Differences Between the Actual and Obtained Percentile Ranks

	5th %ile Equivalents			10th %ile Equivalents		
Skewness	*z* score	*Mdn z* score	*t*-score	*z* score	*Mdn z* score	*t*-score
1.7	+ 7	+ 11	+ 7	+ 6	+ 12	+ 6
1.4	+ 7	+ 12	+ 8	+ 4	+ 10	+ 4
0.6	+ 2	+ 3	+ 3	+ 2	+ 3	+ 2
0	0	1	0	1	- 2	- 1
-0.81	- 2	+ 3	- 2	- 3	- 1	+ 3
-1.42	- 4	- 4	- 4	- 4	- 6	- 3
-1.64	- 3	- 4	- 3	+ 3	- 1	+ 3

Note: Positive numbers reflect overestimation and negative numbers indicate underestimation.

When regarding the difference scores in Table 5.9, it is apparent that the traditional *z* score transformation produces the smallest error out of the three transformations analysed. Whilst Crawford and Howell's (1998) modified *t*-test is designed to accommodate small sample sizes, when using large sample sizes there is no substantial difference between the method and the *z* score when assessing the impact of skewness. Therefore, it is recommended that *z* score transformations be used with normative data of all skewness levels that have large sample sizes (i.e., more than 90). Although the *z* score transformation is deemed adequate, it is also important to acknowledge that this method still introduces an error that will potentially impact the clinical decision making process. Table 5.10 summarises the clinical interpretation when using the 5th percentile as the cut-off for abnormality. Percentiles above the cut-off would, therefore, be interpreted as 'No impairment' while percentiles at or below the 5th percentile would be interpreted as 'impairment'.

Table 5.10.
Clinical Interpretation using 5th Percentile Cut-off for *z* score Transformations

Skewness	*z* score Difference	*Interpretation*
1.7	+ 7	No impairment

1.4	+ 7	No impairment
0.6	+ 2	No impairment
0	0	Impairment
-0.81	- 2	Impairment
-1.42	- 4	Impairment
-1.64	- 3	Impairment

As can be seen, for positively skewed distributions, percentiles that should be reflecting impairment are being misclassified as unimpaired. This raises some ethical and professional issues that need to be addressed in clinical practice.

5.2.1 Implications and Recommendations From Study Three

Clinicians need to avoid error in judgment when using *z* score transformations by making corrections based on the skewness of the normative data they are using. Based on the data in Table 5.10, a regression equation can be used to achieve this result. The *z* score difference using the 5th percentile was plotted with regard to degree of skewness. A 4th order polynomial trend line was fitted to the data (R^2 = 0.985), as depicted in Figure 5.1.

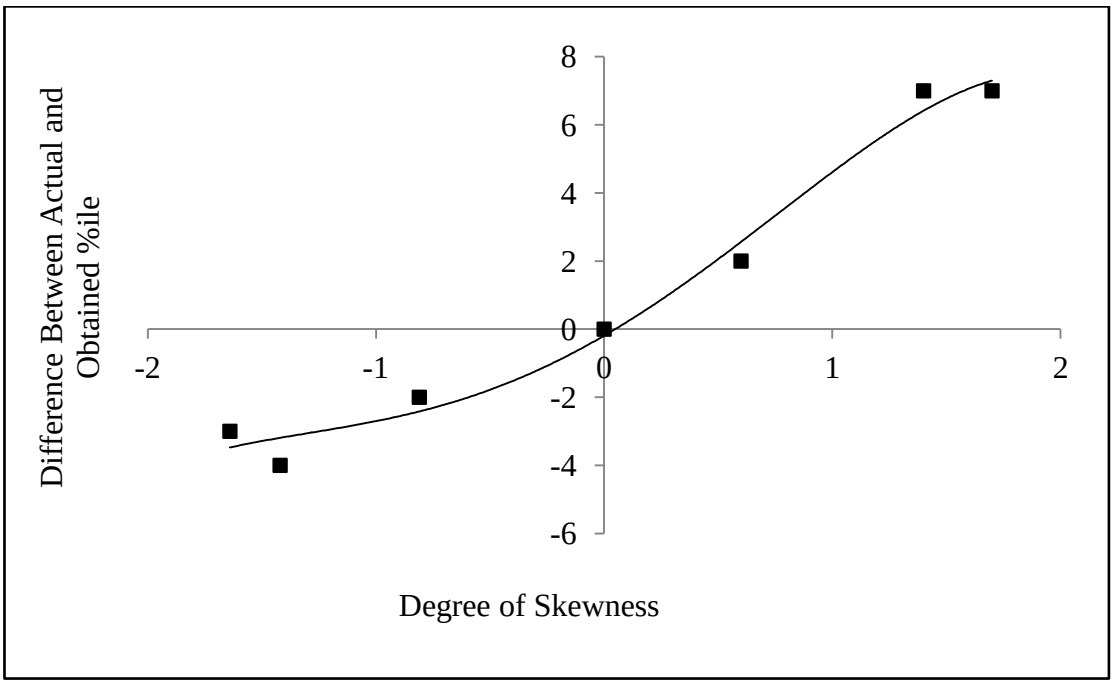

Figure 5.1. The difference between actual and obtained %ile as a function of skewness

This formula accounts for more than 98 percent of the variance in predicting obtained percentile corresponding to the 5th percentile when standardising a score using the *z* score transformation for differing levels of skewness is:

$$y = -0.2299\,x^4 - 0.2433\,x^3 + 1.3765\,x^2 + 3.8999\,x + 0.1918 \qquad \text{Formula 19}$$

Where:

x = the skewness statistic of the sample distribution

Based on these findings, it is recommended that clinicians adhere to the following steps in order to reduce the error introduced when using z score transformations in cognitive assessment:

1. Acquire skewness statistic for the normative data being used.
2. Substitute the skewness statistic into Formula 19.
3. Determine how much correction is needed in order to adjust the obtained percentile (output of Formula 19) into the actual 5th percentile.

For example, in administering TMT B to a client, a clinician can use the skewness statistic to work out what percentile rank represents the 5th percentile or the cut-off for classifying abnormality. Formula 19 indicates that the 5th percentile equivalent for TMT B when using a z score transformation appears as the 12th percentile. Therefore, if a client's score were the 12th percentile when using a *z* score, the clinician would correctly interpret this level as reflecting impairment. If a clinician disregards the effect of skewness on *z* score transformations, the client's performance would be incorrectly classified as not reflecting impairment.

To aid a clinician in performing these corrections when using the 5th percentile cut-off criteria, Table 5.11 can be used. This table displays different levels of skewness and the estimated obtained *z* score produced calculated for each using Formula 19. A clinician can then subtract the corresponding amount in order to correct for skewness. No values have been computed for negative skewness as the resultant z scores indicate percentiles lower than 5 percent and do not pose a risk of being misidentified as being unimpaired.

Table 5.11.

Corrections Required to Reflect 5th Percentile Cut-off as a Function of Skewness

Skewness	Obtained *z* score	Subtract
1.7	12.30	7.30
1.6	12.07	7.07
1.5	11.77	6.77
1.4	11.42	6.42
1.3	11.01	6.01
1.2	10.57	5.57
1.1	10.10	5.10
1.0	9.61	4.61
0.9	9.1	4.1
0.8	8.59	3.59
0.7	8.07	3.07
0.6	7.56	2.56
0.5	7.06	2.06
0.4	6.57	1.57
0.3	6.09	1.09
0.2	5.64	0.64
0.1	5.21	0.21
0	4.81	-

These findings demonstrate that it is necessary to consider skewness when standardising a raw score using normative data. Clinicians should utilise Formula 19 in clinical practice as a means of adjusting the error introduced through skewed distributions. It is, however, important to note that Study Three evaluated this issue using test distributions with large sample sizes. Although the use of optimal sample size is consistent with the findings of Study Two, it fails to address pre-established normative studies with small sample sizes. Therefore, it is necessary to consider the errors associated from each of these three standardisation methods when smaller sample sizes are used. This will be evaluated in Study Four.

5.3 Study Four - Errors with Standardising Skewed Distributions with Small Sample Sizes

This study will evaluate the errors produced when the three standardisation methods outlined in Study Three are applied to raw score distributions with small sample sizes. The rationale for this study is based on the common practice of clinicians to use normative data that have small sample sizes. For example, the norms for the Hooper Visual Organisation Test (i.e., a test of perceptual organisation) developed by Richardson and Marottoli (1996) have sample sizes ranging from 18 to 33 because of data stratification. Although Study Two has demonstrated that the optimal sample size for normative data is a function of skewness, it is understood that many clinicians will continue to use normative data that do not comply with these new findings.

The methodology of this study is similar to Study Three. However, instead of using the raw score distributions of the seven neuropsychological tests with large sample sizes, this study will use only a random sample of 20 cases from each distribution. The descriptive statistics are presented in Table 5.12.

Table 5.12.
Descriptive Statistics of Seven Neuropsychological Tests with N = 20

Test	Skewness	Mean	SD	Median	Median SD
COWAT	0.00	43.30	11.30	43.00	11.30
HVOT	-1.11	25.48	2.35	27.00	2.41
TMT A	1.92	26.92	13.27	22.17	14.14
TMT B	1.04	64.25	26.61	53.77	28.69
Rey 15 Item	-0.96	13.05	2.31	13.5	2.35
WAIS-III Symbol Search	-0.90	29.8	11.6	32.00	11.81
WTAR	-0.90	36.95	7.38	37.5	7.40

Note: Median SD was calculated using Formula 18

It is important to note that the skewness statistics for the tests differ from the skewness statistics shown in Study two. This was due to the small sample size creating unstable skewness statistics.

The first step in this study was to analyse each distribution and find the raw scores corresponding to different percentiles in the distributions. The percentiles chosen included the 1st, 5th, 10th, 25th, 50th, 75th, 90th, and 95th. Raw scores corresponding to the 99th percentile for each distribution were unable to be calculated due to the small sample size. Neuropsychological assessment, which focuses on abnormally low scores, means that the omission of the 99th percentile is not of concern. The results are presented in Table 5.13 below.

Table 5.13.
Corresponding Raw Scores for Different Percentiles in the Underlying Distribution Based on a Sample Size of 20

	%iles								
Test	1	5	10	25	50	75	90	95	99
COWAT	22.00	22.30	28.40	33.25	43.00	52.50	60.00	60.95	-
HVOT	20.50	20.60	22.60	25.63	27.00	27.88	29.45	29.5	-
TMT A	15.18	15.18	15.37	19.39	22.17	28	58.62	60.95	-
TMT B	32.30	32.38	34.61	45.76	53.77	83.25	113.55	126.44	-
WAIS-III Symbol Search	0.00	0.65	13.20	21.75	32.00	38.00	42.70	48.70	-
WTAR	21.00	21.35	28.10	30.50	37.50	43.00	45.90	48.85	-
Rey 15 Item	8.00	8.00	8.30	12.00	13.50	15.00	15.00	15.00	-

Using the known raw scores for each percentile, the three standardisation methods could then be applied and compared for each distribution. For each method, the *z* score or *t*-score was calculated and was converted into the corresponding percentile rank using Table 5.12. This percentile rank was then compared to the original percentile rank, and the difference between them was calculated. Tables 5.14 through to 5.20 present the data for each of the seven tests. It is important to note that the 5^{th} and 10^{th} percentiles are most important for inferring impairment in neuropsychological assessment.

Table 5.14.
COWAT – Comparisons of Three Standardisation Methods with N = 20

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-1.88	3	2	-1.86	4	3	-1.93	3	2
5	-1.86	4	-1	-1.83	4	-1	-1.90	3	-2
10	-1.32	10	0	-1.29	10	0	-1.35	9	-1
25	-0.90	19	-6	-0.86	20	-5	-0.91	19	-6
50	-0.02	49	-1	0	50	0	-0.02	49	-1
75	0.81	80	5	0.84	80	5	0.83	80	5
90	1.48	94	4	1.50	94	4	1.51	94	4
95	1.56	95	0	1.59	95	0	1.60	95	0

Table 5.15.
HVOT – Comparisons of Three Standardisation Methods with n = 20

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-3.69	0.03	-0.97	-4.23	0.03	-0.97	-3.78	0.03	-0.97
5	-1.91	3	-2	-2.50	1	-4	-1.95	3	-2
10	-0.84	20	10	-1.45	8	-2	-0.86	20	10
25	0.01	51	26	-0.62	27	2	0.01	51	26
50	0.65	75	25	0	50	0	0.66	75	25
75	1.07	86	11	0.41	67	-8	1.10	87	12
90	1.50	94	4	0.83	80	-10	1.53	94	4
95	1.71	96	1	1.04	86	-9	1.75	96	1

Table 5.16.
TMT A – Comparisons of Three Standardisation Methods with n = 20

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-0.88	19	18	-0.49	32	31	-0.91	19	18
5	-0.88	19	14	-0.49	32	27	-0.91	19	14
10	-0.87	19	9	-0.48	32	22	-0.89	19	9
25	-0.57	29	4	-0.20	43	18	-0.58	28	3
50	-0.36	36	-14	0	50	0	-0.37	36	-14
75	0.08	54	-21	0.41	67	-8	0.08	47	-28
90	2.39	99	8	2.58	99.1	9.1	2.45	99	9
95	2.56	99.1	4.1	2.74	99.4	4.4	2.63	99.2	4.2

Table 5.17.
TMT B – Comparisons of Three Standardisation Methods with n = 20

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-1.20	12	11	-0.75	23	22	-1.23	11	10
5	-1.20	13	8	-0.75	23	18	-1.22	11	6
10	-1.11	14	4	-0.67	25	15	-1.14	13	3
25	-0.69	25	0	-0.28	39	14	-0.73	24	1
50	-0.39	35	-35	0	50	0	-0.40	34	16
75	0.71	76	1	1.03	85	10	0.73	77	2
90	1.85	97	7	2.08	98	8	1.90	97	7
95	2.34	99	4	2.53	99	4	2.39	99	4

Table 5.18.
Rey 15 Item – Comparisons of Three Standardisation Methods with n = 20

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-2.19	2	1	-2.34	1	0	-2.24	1	0
5	-2.19	2	-3	-2.34	1	-4	-2.24	1	-4
10	-2.06	2	-8	-2.21	1	-9	-2.11	2	-8
25	-0.45	33	8	-0.64	26	1	-0.47	32	7
50	0.19	58	8	0	50	0	0.20	58	8
75	0.84	80	5	0.64	74	-1	0.87	81	6
90	0.84	80	-10	0.64	74	-16	0.87	81	-9
95	0.84	80	-15	0.64	74	-21	0.87	81	-14

Table 5.19
WAIS-III Symbol Search – Comparisons of Three Standardisation Methods n = 20

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-2.57	1	0	-2.71	0.8	-0.2	-2.63	0.9	-0.1
5	-2.51	1	-4	-2.65	0.9	-4.1	-2.57	1	-4
10	-1.43	8	-2	-1.59	6	-4	-1.47	7	-3
25	-0.69	25	0	-0.87	19	-6	-0.71	24	-1
50	0.19	43	-7	0	50	0	0.19	58	8
75	0.71	76	1	0.51	70	-5	0.72	77	2
90	1.11	87	-3	0.91	82	-8	1.14	88	-2
95	1.63	95	0	1.41	93	-2	1.67	95	0

Table 5.20.
WTAR – Comparisons of Three Standardisation Methods with N = 20

Actual %ile	*z* – score	%ile	Difference	*Mdn* *z* – score	%ile	Difference	Crawford & Howell's (1998) *t* - score	%ile	Difference
1	-2.16	2	1	-2.23	1	0	-2.21	1	0
5	-2.11	2	-3	-2.18	2	-3	-2.12	2	-3
10	-1.20	12	2	-1.27	10	0	-1.20	12	2
25	-0.87	19	-6	-0.95	17	-8	-0.88	19	-6
50	0.07	53	3	0	50	0	0.07	53	3
75	0.82	80	5	0.74	78	3	0.82	80	5
90	1.21	89	-1	1.14	88	-2	1.21	89	-1
95	1.61	95	0	1.53	94	-1	1.61	95	0

The difference scores for the 5th and 10th percentiles were then summarised for each of the seven tests, as presented in Table 5.21.

Table 5.21.
Summary of Differences Between Obtained and Actual Percentiles

		5th %ile equiv			10th %ile equiv		
Test	Skew	*z score*	*Mdn z score*	*t-score*	*Z* score	*Mdn z* score	*t-* score
TMT A	1.92	+ 14	+ 27	+ 14	+ 9	+ 22	+ 9
TMT B	1.04	+ 8	+ 18	+ 6	+ 4	+ 15	+ 3
COWAT	0.00	- 1	- 1	- 2	0	0	- 1
WTAR	-0.90	- 3	- 3	- 3	+ 2	0	+ 2
WAIS-III SS	-0.90	- 4	- 4.1	- 4	- 2	- 3	- 3
Rey 15 Item	-0.96	- 3	- 4	- 4	- 8	- 9	- 8
HVOT	-1.11	- 2	- 4	- 2	+ 10	- 2	+ 10

Note: Positive numbers reflect overestimation and negative numbers indicate underestimation.

As can be appreciated from the above table, the traditional *z* score transformation and Crawford and Howell's (1998) modified *t*-score produce very similar results. Consistent with the results of Study Three, it is recommended that the *z* score transformation be used with normative data of all skewness levels regardless of sample size. It should be noted that this is counter to the findings and recommendations of Crawford et al., (2006). In their study, the modified *t*-test produced fewer errors than the *z* score transformation and was the recommended method when using small sample sizes. Regardless, it is crucial to evaluate how much error is introduced into clinical decision making when using the *z* score transformation on small sample sizes. Table 5.22 presents the clinical interpretation of the obtained *z* scores if a clinician was to use the 5th percentile cut-off on small sample sizes compared with those with adequate sample sizes.

Table 5.22.
Clinical Interpretation using 5th Percentile Cut-off for z score Transformations using Different Sample Sizes

Skewness	Adequate *n*	*n* = 20
>1.7	+ 7	+ 14

>1.0	+ 7	+ 8
0.6	+ 2	- 1
0	0	- 4
-0.81	- 2	- 3
-1.42	- 4	- 3
-1.64	- 3	- 2

This new information can then be superimposed into the graph from Study Three (Figure 5.1) and is depicted in Figure 5.2.

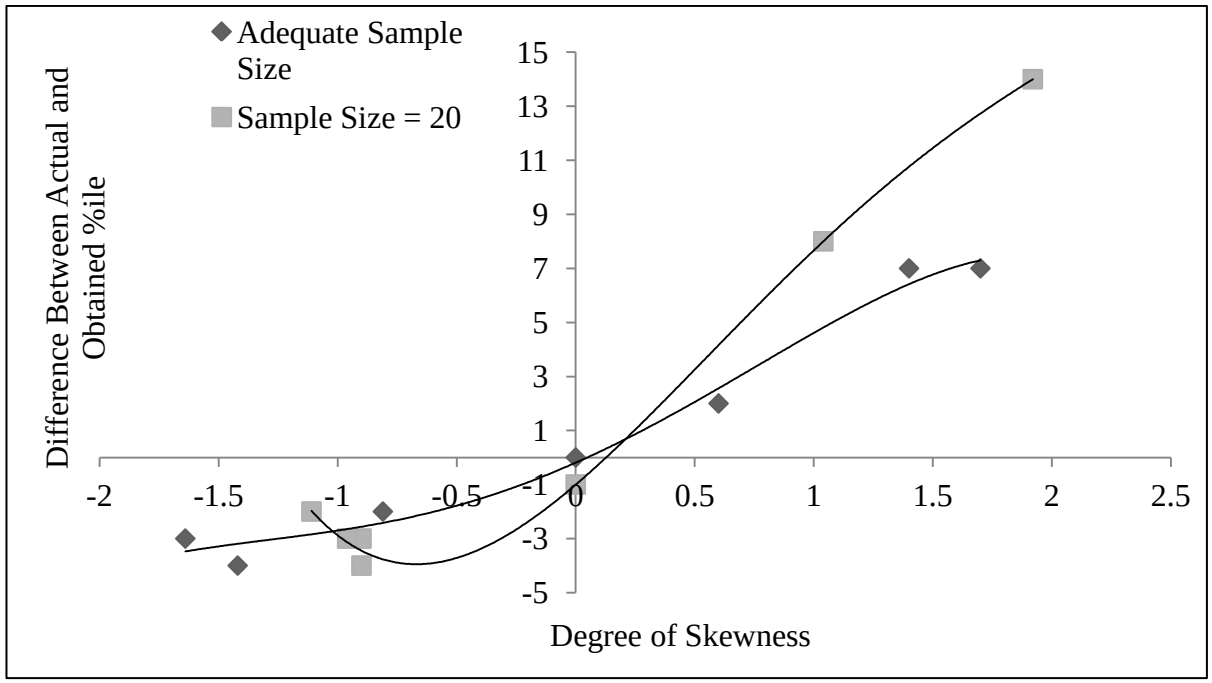

Figure 5.2. The difference between actual and obtained percentile as a function of skewness and sample size.

As can be appreciated, the consequences of standardisation using inadequate normative data with positively skewed distributions are extensive. For distributions with skewness statistics greater than 1.7, the error between the actual and obtained percentiles using the *z* score transformation is approximately doubled. It is, therefore, recommended that clinicians should discontinue using normative samples with inadequate sample sizes in their clinical decision making processes. From Study Two, the minimum sample size required for stable means and standard deviations for highly negatively skewed distributions (e.g., HVOT) was 30 (Table 3.12). Therefore clinicians should no longer use sample sizes with less than 30 in each cell with these types of distributions.

5.4 Summary

The issue of skewed distributions has particular importance to clinicians because of the errors introduced during standardisation. The study by Donnell et al. (2011) demonstrated that the errors are greater for linear-transformed scores than their normalised-standard score counterparts in skewed distributions. Although these findings are important, the fact remains that some clinicians still utilise linear transformations in clinical practice regardless or without considering the results of Donnell et al. (2011). Study Three, therefore aimed to assess how much error is introduced when different linear transformations are used on raw-score distributions that vary in levels of skewness. Results indicated that the traditional and commonly used *z* score transformation was the best method to be used for skewed distributions. However, the error rates are still of considerable concern for positively skewed distributions. A regression formula was created in order to perform corrections when using the 5^{th} percentile cut-off for abnormality.

Study Four investigated the effects on clinical interpretation if a clinician continues to transform scores using skewed normative data with inadequate sample sizes. Results indicated that the error rates associated with positively skewed distributions are even larger when using small sample sizes. It was, therefore, recommended that clinicians discontinue using normative samples with sample sizes below those recommended in Study Two.

CHAPTER SIX
GENERAL DISCUSSION, CONCLUSIONS AND IMPLICATIONS

6.1 Overview

In neuropsychological assessment little can be understood without comparing an individual's raw score with a normative sample. This is achieved through standardisation – converting a raw score to a standard score and thereby creating a basis for comparison. While there is a wealth of literature on reliability and validity, far less consideration is given to the psychometrics related to the standardisation process. One paramount issue is related to the sample size of the normative sample. The current lore indicates a minimum sample size of 50 is adequate, however this number appears to be based on a limited psychometric base and does not reflect considerations of skewed distributions commonly seen in neuropsychology.

The task of psychological assessment is complex with potential errors negatively influencing clinical interpretation. Currently, many clinicians are unlikely to consider the error made during the standardisation process, essentially accepting whatever normative data that is available as adequate. As a result, the primary goal of the current thesis was to evaluate the psychometric literature related to standardisation concepts and processes required in clinical neuropsychology. In addition, the thesis also aimed to investigate the current understanding of sample sizes in normative samples, how this is influenced by different distributions, and the potential errors involved in the decision making process. Guidelines for determining adequate sample sizes were developed for normal and skewed distributions and recommendations provided to ensure better evidence-based practice for clinicians.

This chapter will present a summary of each of the studies undertaken in this thesis and will specifically discuss the implications and applications for clinical practice in each case. In addition, this chapter will address the limitations of the current research and indicate directions for future research.

6.2 General Discussion and Conclusion of Results

6.2.1 Summary, Recommendations and Implications of Study One

After reviewing the current literature, it was found that the issue regarding minimum sample sizes in normative studies is often arbitrary. The consensus is that sample sizes should have a minimum of 50 subjects in order to be deemed appropriate for use. Whilst most normative studies have sample sizes in excess of this figure, the difficulty is that the data are typically stratified by demographic variables such as age, education, and gender. This ultimately means that sample sizes for each of these stratified cells are commonly less than 50. In some cases, frequently used normative data have sample sizes with $n = 20$ in each cell. Study One aimed to empirically evaluate the issue of sample size in normative studies. Results indicated that the optimal *n*'s required to produce stable measures of central tendency are dependent on the level of skewness of the raw score data. Interestingly, normally distributed data require 70 cases in each cell.

A formula was developed to aid clinicians in determining the minimum *n* for different levels of skewness. However as a general rule, if skewness statistics are not provided then it is recommended that clinicians use only normative data with a minimum of $n = 90$ in each cell. Clinicians and researchers wishing to undertake normative studies designed to generate stable measures of central tendency and variance are advised to adhere to the following steps:

- Step 1: Acquire a minimum of 30 cases for the normative study. A sample size of 30 was the smallest number required for stables means and standard deviations for highly negatively skewed distributions (i.e., Table 3.12). If the data is to be stratified (e.g., by age and education), then each cell must have a minimum of 30 cases.
- Step 2: Analyse this data and generate a skewness statistic.
- Step 3: Use Formula 11 to compute the minimum *n* required in each cell for a stable mean and standard deviation at the indicated level of skewness.
- Step 4: If the computed *n* is equal to cases already obtained, then the minimal sample size requirement is satisfied. If a larger minimum sample size is indicated than the cases already obtained, then more cases will be needed until the minimum in each cell is satisfied.

The implication of Study Two is that many established normative studies currently have inadequate sample sizes. Using the general rule outlined above (i.e., only use normative studies with $n > 90$ when skewness statistics are not provided), the normative data of two neuropsychological tests contained within the Handbook of Normative Data for Neuropsychological Assessment (Mitrushina et al., 2005) were analysed. This book contains arguably the most comprehensive listings of normative data made available to clinicians. The tests were the Trail Making Test (TMT), and the Color Trails Test (CTT).

In total, 49 normative studies (including sub-studies) for the TMT are included in the Handbook of Normative Data for Neuropsychological Assessment (Mitrushina et al., 2005). Out of the 49 normative studies, 37 had inadequate sample sizes in one or more of their stratified cells. In total, out of all the 243 stratified cells across the 47 normative studies, 71% had inadequate sample sizes.

For the CTT, 10 normative studies (including sub-studies) were included in the Handbook. All of the 10 normative studies for the CTT had at least one cell with inadequate sample sizes and of the 53 cells, 87% had sample sizes less than 90.

Although these are only two examples, it raises the question of the adequacy of established neuropsychological normative data. The astute reader will note an interesting implication with regards to the TMT samples sizes used in the worked example in Chapter two. When referring to Table 3.12, TMT A requires a minimal sample of 80 and TMT B requires a minimal sample size of 60 based on the skewness statistics. However when the normative data was stratified, TMT A sample sizes ranged from 39 to 251 and TMT B sample sizes ranged from 38 to 250. It should be noted however that the spread of the normative data used in this example likely reflects the consequences of the sampling method. What this does highlight is the normative data collected for the TMT utilised in this thesis has extensive oversampling in one of the cells (i.e, <50 years old, more than 12 years of education) and under-sampling in another (i.e., more than 50 years of age and less than 12 years of education). Therefore further sampling will need to be undertaken in the cells that have inadequate sample sizes before publication of this data for normative purposes would be recommended.

The overall implication is that further research needs to be conducted to update existing normative studies with inadequate sample sizes. This should be completed with due consideration of the skewness of the raw score data. Furthermore, skewness statistics should be provided as a standard psychometric characteristic for all existing and future normative studies.

6.2.2 Summary and Recommendations of Study Two

Study Two was concerned with the error introduced when linear transformations are used on different skewed distributions. This issue is paramount, as the research has demonstrated that there is an increase in errors when using a linear transformation on a skewed distribution. This is because, unlike a normal distribution, the mean and the median are not the same in skewed distributions. Regardless of this research, clinicians still adopt the practice of conducting z score transformations presumably due to the ubiquitous provision of sample means and standard deviations in all normative studies. It is, therefore, necessary to empirically evaluate the impact this is having on clinical decision-making. Using three different standardisation methods on seven neuropsychological test distributions with adequate sample sizes (i.e., n = 90+), the results indicated that the traditional *z* score transformation was the most effective and introduced less errors in judgement. While this may allay concerns regarding conventional practice, albeit through ignorance rather than evidence-based, the errors produced were still of concern, particularly for highly positively skewed distributions. For example, in the extremely positively skewed distribution of the TMT B, the 12th percentile obtained using the *z* score transformation in fact reflects the true 5th percentile. This is a seven-point overestimation. Furthermore, in clinical practice, this particular score could be classified as unimpaired when in fact it is reflective of impairment (using the 5th percentile as a cut-score). A regression formula was created that helps to correct data with adequate sample sizes for skewness. Corrections can then be applied to the cut-score percentile level in order to account for the overestimation.

An important finding from this study is that skewness statistics greater than +1.0 produce 5th percentile equivalents that are above the 10th percentile. As a general rule, when standardising scores from tests that have skewness statistics greater than 1.0, Formula 19 should be used and corrections made. Secondly, when standardising scores from tests that have skewness statistics less than +1.0 but more than 0, the general rule should be to use a 10th percentile cut-off score for classifying impairment. This will ensure skewness is considered in the interpretation process. For tests with skewness statistics less than 0, the fifth percentile cut-off score is adequate. Table 6.1 summarises these general rules.

Table 6.1.
General Rules for z score Transformations as a Function of Skewness

General Rule	Skewness Range of Normative Data	Cut-score for Impairment
1	1.0+	Caution. Consult Formula 11
2	0.0 – 1.0	Use 10th percentile
3	< 0.0	Use 5th percentile

6.2.3 Summary and Recommendations of Study Three

Study three was concerned with the errors produced if clinicians choose to ignore the findings from Study One and continue to use normative data with small sample sizes. The three standardisation methods used in Study Two were applied to

sample sizes of n = 20 on various skewed raw score distributions. Results indicated the error in judgement is nearly doubled on positively skewed distributions. Therefore, if a clinician chooses to use inadequate sample sizes, producing unstable measures of central tendency and skewness statistics, the problem is potentially severe. It is recommended that normative data with sample sizes less than 30 should not be used in clinical practice as this number represents the minimum sample size needed to produce stable means and standard deviations in highly negatively skewed distributions (i.e., HVOT).

6.3 General Recommendations

Overall, the purpose of this dissertation was to illustrate and evaluate important psychometric issues relevant to the standardisation process. The results of the three empirical studies are two-fold. Firstly, they outline the importance of adequate sample size as a function of skewness. Secondly, they demonstrate the appropriate procedures for transforming raw scores into standardised scores while adjusting for errors introduced through level of skewness. Table 6.2, which combines these results, is intended to provide a valuable resource for clinicians by presenting the optimum sample size for different levels of skewness (i.e., derived using Formula 11) and an approximation of the judgement errors and corrections required as calculated using derived formula 19.

Table 6.2.
Optimum Sample Size and Estimated Judgement Errors for Differing Skewness Levels at the 5th percentile

Skewness	Minimum *n*	*z* score percentile	Subtract
1.7	61	12.30	7.30
1.6	69	12.07	7.07
1.5	75	11.77	6.77
1.4	79	11.42	6.42
1.3	82	11.01	6.01
1.2	84	10.57	5.57
1.1	85	10.10	5.10
1.0	86	9.61	4.61
0.9	85	9.1	4.1
0.8	84	8.59	3.59
0.7	83	8.07	3.07
0.6	81	7.56	2.56
0.5	80	7.06	2.06
0.4	78	6.57	1.57
0.3	76	6.09	1.09
0.2	74	5.64	0.64
0.1	72	5.21	0.21
0	70	4.81	-
-0.1	69	-	-
-0.2	67	-	-
-0.3	66	-	-
-0.4	64	-	-
-0.5	63	-	-
-0.6	62	-	-
-0.7	61	-	-

-0.8	59	-	-
-0.9	58	-	-
-1.0	56	-	-
-1.1	54	-	-
-1.2	51	-	-
-1.3	48	-	-
-1.4	43	-	-
-1.5	38	-	-
-1.6	32	-	-

Note: Skewness levels less than 0 do not require a correction.

As can be appreciated by the above table, the requirement underlying the correct use of this resource is the skewness statistic. The dilemma facing clinicians is how to evaluate established normative studies where skewness statistics are not made available. This will be a common problem faced by clinicians. For example, in the Handbook of Normative Data for Neuropsychological Assessments (Mitrushina et al., 2005), no skewness statistics are provided for any of the normative studies. Unfortunately, skewness statistics cannot be calculated without access to the raw score data. In order to aid clinicians, Table 6.3 provides skewness statistics for raw score distributions of all neuropsychological tests from the three databases used throughout this thesis. Also included in Table 6.3 is the estimated optimal *n* for each and the rule from Table 6.1 above that should be applied when standardising the particular test.

By way of summary, Table 6.4 presents levels of consideration by practitioners in relation to skewness, sample sizes, and the use of normative data. It is intended as a way for clinicians to evaluate the extent to which they are prepared to alter their practices to reduce error in relation to these issues and how best to act.

Table 6.3.
Skewness Statistics for 45 Neuropsychological Tests and Calculated z score Equivalents Based on the 5th Percentile Cut-off Score (Presented in Order of Skewness).

Test	Database*	*N*	Skewness	*SE*	Optimum *n*	Rule (Table 6.1)
TMT B	1	508	1.65	0.11	60	1
TMT A	1	508	1.40	0.11	79	1
RAVLT – Interference	1	200	1.25	0.17	84	1
Stroop Colour-Word	1	729	0.70	0.91	83	2
RAVLT – Trial I	1	200	0.63	0.17	82	2
COWAT	1	935	0.60	0.08	82	2
WAIS-III Digit Span Backwards	2	1250	0.54	0.07	80	2
Visual Form Discrimination	1	345	0.41	0.13	78	2
WAIS-III Digit Span	2	1250	0.35	0.07	77	2
Symbol Digit Modalities Test – Oral	1	629	0.20	0.10	74	2
WAIS-III Block Design	2	1250	0.18	0.07	74	2
Rey Visual Design Learning Test – Trial I	1	166	0.12	0.19	72	2
WAIS-III Arithmetic	2	1250	0.10	0.07	72	2
Stroop Word	1	729	0.09	0.91	72	2
WAIS-III Matrix Reasoning	2	1250	0.05	0.07	71	2
WAIS-III Information	2	1250	0.05	0.07	71	2
WAIS-III Symbol Search	2	1250	0.00	0.07	70	3
Stroop Colour	1	729	-0.01	0.91	70	3
Symbol Digit Modalities Test – Word	1	628	-0.07	0.10	69	3
WAIS-III Digit Symbol	2	1250	-0.08	0.07	69	3
Speed of Comprehension	1	787	-0.11	0.09	69	3
WAIS-III Vocabulary	2	1250	-0.17	0.07	68	3
WAIS-III Letter-Number Sequencing	2	1250	-0.18	0.07	67	3
WAIS-III Picture Arrangement	2	1250	-0.32	0.07	65	3
Rey Visual Design Learning Test – Trial	1	166	-0.33	0.19	65	3

I-IV						
WAIS-III Similarities	2	1250	-0.39	0.07	65	3
National Adult Reading Test (NART)	1	160	-0.40	0.19	64	3
WAIS-III Comprehension	2	1250	-0.43	0.07	64	3
RAVLT – Trial I-IV	1	200	-0.48	0.17	63	3
Rey Visual Design Learning Test – Delayed	1	166	-0.50	0.19	63	3
Test of Premorbid Functioning (TOPF)	1	145	-0.51	0.20	63	3
Conceptual Level Analogies Test (CLAT)	1	265	-0.54	0.15	63	3
RAVLT – Trial VI	1	200	-0.65	0.17	61	3
Spot the Word	1	783	-0.70	0.08	61	3
Rey Visual Design Learning Test –Trial I	1	166	-0.76	0.19	60	3
Wechsler Test of Adult Reading (WTAR)	1	389	-0.83	0.12	59	3
RAVLT – Trial VII	1	200	-0.83	0.17	59	3
Shipley Institute of Living Scale (Shipley)	1	404	-0.92	0.12	57	3
Boston Naming Test (BNT)	1	571	-1.07	0.10	54	3
WAIS-III Picture Completion	2	1250	-1.09	0.07	54	3
Judgement of Line Orientation Test (JLO)	1	379	-1.31	0.13	47	3
Rey 15 Item Test	3	272	-1.42	0.15	42	3
Hooper Visual Organisation Test (HVOT)	1	379	-1.64	0.13	30	3

Note: Database 1 = USQ Normative data, 2 = WAIS-III/WMS-III; Standardisation data, 3 = Clinical data

Table 6.4.
Levels of Consideration for Clinicians in relation to Skewness, Sample Size, and Standardisation

Level	Considerations and Recommendations
0	Avoidance. Utilise established normative data only where there is a minimum of n = 90 in each cell. This will ensure that the data have stable means and standard deviations. Do not use tests that have extreme positive skewness. Clinicians at this level seek to remove the need for consideration of sample and distribution factors by choosing only those measures that do not suffer from them.
1	Assimilation. Utilise the skewness estimates for a range of neuropsychological tests presented in Table 6.3 to inform clinical decision-making and utilise the percentile heuristic to adjust cut-offs for abnormality accordingly. At this level the clinician attempts to correct for skewness and inadequate sample sizes within their existing test battery.
2	Accommodation. At this level, clinicians actively engage in the process of determining the influence of sample size and skewness on their test battery, through examining their own data and modifying test selection procedures just as they do for validity and reliability considerations in designing a test battery that accommodates psychometric issues. That is, clinicians should only select tests for their battery if the accompanying norms satisfy the sample size requirements. This would require the established normative data to include skewness statistics so sample size can be evaluated and appropriate standardisation methods can be employed. Clinicians could also consult Table 6.2 to calculate error associated with level of skewness and consequently correct their clinical interpretation.

6.4 Limitations and Future Directions

The lack of skewness statistics made available by test publishers and researchers presents a primary barrier to the application of the recommendations presented in this dissertation. Without skewness statistics, clinicians are forced to use the basic level of psychometric consideration or Level One of Table 6.4 (i.e., Avoidance). However, with inclusions of skewness statistics in manuals and/or papers, clinicians can begin to apply the full statistical algorithm and evaluate the adequacy and potentials errors themselves.

Another limitation is that for this thesis, all that has been analysed are the tests used in the normative studies undertaken in Australia and the WAIS-III/WMS-III standardisation and education oversample. There is further need to analyse more tests and normative databases to evaluate the extent to which skewness coefficients may change with different populations. Although there is awareness to use reliable and stable measures, coefficients are still treated as constants rather than being treated as variable measures which have confidence intervals in their own right. Caution therefore needs to be applied to the skewness statistics provided in Table 6.3. These are skewness statistics provided for the normative data analysed in this research and should not be considered a constant or fixed number. Further research therefore needs to evaluate whether the skewness statistics found here are similar to those found in other normative data, evaluating whether skewness is a characteristic of a test that overrides or persists in different samples, countries, cultures, and/or languages.

6.5 Conclusion

In current practice, clinicians are still using and relying on extremely poor normative data when standardising test scores. The fault is not with the clinicians but with the limited availability of adequate data. The current research aims to bring awareness to the instability and potential of making interpretation errors when using the current normative data for a large range of neuropsychological tests. Clinicians now have guidelines that indicate whether a sample size is adequate based on the level of skewness of the distribution and whether this skewness will introduce errors during raw score transformations. In particular, this research has demonstrated that the influence of skewness is important during the standardisation process, especially with positively skewed data. While negatively skewed distributions skew the percentile, they do so in a direction that does not increase the likelihood of an incorrect interpretation. Positively skewed data, however, overestimate the percentile ranks with the implication that clinicians are interpreting scores reflecting impairment as unimpaired. The major finding of the current research is that the level of error introduced when transforming data using normative samples of less than 30 is significant and therefore should not be used in clinical practice. Thirty is the minimum sample size found in Study Two for highly negatively skewed distributions (see Table 3.12). This is an important finding given the current use of such small sample sizes in clinical neuropsychology. It is therefore hoped that this research will contribute to the efforts of clinicians who seek to reduce error in their decision-making and to improve the standards used in neuropsychology.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing international edition* (7th ed.). Sydney: Prentice Hall.

American Psychological Association (2010). *Ethical Principles of Psychologists and Code of Conduct.* Washington, DC: Author

Anderson, V. A., Lajoie, G. (1996). Development of memory and learning skills in school-aged children: A neuropsychological perspective. *Applied Neuropsychology, 3/4.* 128-139.

Ardila, A. (1995). Directions of research in cross-cultural neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17*(1), 143-150.

Army Individual Test Batter (1944). *Manual of directions and scoring.* Washington DC: War Department, Adjutant General's Office.

Australian Bureau of Statistics. (2000). *Australian social trends.* Canberra, Australia: Author.

Australian Bureau of Statistics. (2011). *Census of Population and Housing: Counts of Aboriginal and Torres Strait Islander Australians* (cat. no. 2901.0). Retrieved from http://www.abs.gov.au

Australian Psychological Society (2007). *Code of Ethics.* Victoria: Australian Psychological Society Limited

Barton, J. J. S., Cherkasova, M. V., Press, D. Z., Intriligator, J. M., & O'Connor, M. (2003). Developmental prosopagnosia: A study of three patients. *Brain and Cognition, 51*(1), 12-30.

Beaumont, J. G. (2008). *Introduction to Neuropsychology* (2nd ed.). New York, NY: The Guilford Press.

Benton, A. L., Hamsher, K., & Sivan, A. B. (1983). *Multilingual aphasia examination* (3rd ed.). Iowa City, IA: AJA Associates.

Benton, A. L., Varney, N. R., & Hamsher, K. (1978). Visuospatial judgement: A clinical test. *Archives of Neurology, 35,* 364-367.

Boake, C. (2000). Edouard Claparede and the auditory verbal learning test. *J Clin Exp Neuropsychol, 22*(2), 286-292.

Bornstein, R. A., & Suga, L. J. (1988). Educational level and neuropsychological test performance in healthy elderly subjects. *Developmental Neuropsychology, 4,* 17-22.

Bowman, M. L. (2002). The perfidy of percentiles. *Archives of Clinical Neuropsychology, 17,* 295-303.

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology, 13,* 528-538.

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology, 50* (3), 196-209.

Budescu, D. V. (1987). Selecting the equating method: Linear vs. equipercentile. *Journal of Educational Statistics, 12,* 33-43.

Bulmer, M. G. (1979). *Principles of Statistics.* New York: Dover.

Campbell, A. L., Jr, Ocampo, C., DeShawn, R. K., Lewis, S., Combs, S., Ford-Booker, P., … Hastings, A. (2002). Caveats in the neuropsychological assessment of African Americans. *Journal of the National Medical Association, 94*(7), 591-601.

Canadian Psychological Association (2000). *Canadian Code of Ethics for Psychologists* (3rd ed.). Ottawa: Ontario: Author.

Capitani, E., & Laiacona, M. (2000). Classification and modelling in neuropsychology: From groups to single case. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology: vol.1* (2nd ed., pp. 53-76. Amsterdam: Elsevier.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21*(4). 559-566.

Charter, R. A. (2001). Damn the precision, full speed ahead with the clinical interpretation. *Journal of Clinical and Experimental Neuropsychology, 23*(5), 692-694.

Cicchetti, D. V. (1981). Guidelines, criteria, and rules of thumb for evaluating normed and standardised assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia, 40*(8), 1196-1208.

Crawford, J. R., & Garthwaite, P. H. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology, 19,* 318-331.

Crawford, J. R., & Garthwaite, P. H. (2008). On the optimal sample size for normative samples in neuropsychology: Capturing the uncertainty when normative data are used to quantify the standing of a neuropsychological test score. *Clinical Neuropsychology, 14*, 99-117.

Crawford, J. R., & Garthwaite, P. H. (2009). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *The Clinical Neuropsychologist, 23*, 193-204.

Crawford, J. R., & Garthwaite, P. H. (2012). Single-case research in neuropsychology: A comparison of five forms of t-test for comparing a case to controls. *Cortex, 48,* 1009-1016.

Crawford, J. R., Garthwaite, P. H., Azzalini, A., Howell, D. C., & Laws, K. R. (2006). Testing for a deficit in single-case studies: Effects of departures from normality. *Neuropsychologia, 44*, 666-677.

Crawford, J. R., Garthwaite, P. H., & Gault, C. B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardised neuropsychological test batteries: A generic method with applications. *Neuropsychology, 21,* 419-430.

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist, 23*, 1173-1195.

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small sample. *The Clinical Neuropsychologist, 12*, 482-486.

Davis, H.C. (1857). *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections: A Translation of Gauss's "Theoria Motus".* Boston: Little, Brown and Co.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *The California Verbal Learning Test: Adult version manual.* San Antonio, TX: Psychological

Corporation.
Deming, W. E., & Stephan, F. F. (1940). On a least squared adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics, 11,* 427-444.
Donnell, A. J., Belanger, H. G., & Vanderploeg, R. D. (2011). Implications of psychometric measurement for neuropsychological interpretation. *The Clinical Neuropsychologist, 25* (7), 1097-1118.
Dotson, V. M., Kitner-Triolo, M., Evans, M. K., & Zonderman, A. B. (2008). Literacy-based normative data for low socioeconomic status African Americans. *The Clinical Neuropsychologist, 22*, 989-1017.
Field, A. (2009). *Discovering Statistics using SPSS* (3rd ed.). London: Sage Publications Ltd.
Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.
Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.
Flynn, J. R. (1994). IQ gains over time. In R. J. Sternberg (Ed.), *The encyclopedia of human intelligence* (pp. 617–623). New York: Macmillan.
Flynn, J. R. (1998a). WAIS-III and WISC-III IQ gains in the United States from 1972 to 1995: How to compensate for obsolete norms. *Perceptual and Motor Skills, 86,* 1231–1239.
Flynn, J. R. (1998b). IQ Gains Over Time: Toward Finding the Causes. In U. Neisser (Ed.), *The Rising Curve: Long-Term Gains in IQ and Related Measures* (pp. 25-66). Washington, DC: American Psychological Association.
Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54,* 5–20. Retrieved from: http://acdlaonline.com/zoomdocs/presentations
Forrester, G., & Geffen, G. (1991). Performance measures of 7- to 15-year-old children on the auditory verbal learning test. *The Clinical Neuropsychologist, 5,* 345-359.
Goldstein, E. B. (2005). *Cognitive Psychology.* Belmont, CA: Thomson Wadsworth.
Gregory, R. J. (2007). *Psychological testing: History, principles, and applications* (5th ed.). Boston, MA: Pearson Education.
Guilford, J.P. (1936). *Psychometric Methods.* New York: McGraw-Hill.
Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2008). Adjusting IQ scores for the flynn effect: Consistent with the standard of practice? *Professional Psychology: Research and Practice, 39,* 619-625.
Hooper, H. E. (1948). *A study in the construction and preliminary standardisation of a visual organisation test for use in the measurement of organic deterioration.* Unpublished master's thesis, University of Southern California, Los Angeles.
Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology, 10,* 120-124.
Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston Naming Test (Revised 60-item version).* Philadelphia: Lea & Febiger.
Kaplan, R. M., & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth
Kramer, J. H. (1990) Guidelines for interpreting WAIS-R subtest scores. *Psychological Assessment, 2* (2), 202-205.

Ley, P. (1972). *Quantitative aspects of psychological assessment.* London: Duckworth.

Lezak, M. D. (1976) *Neuropsychological Assessment*. New York: Oxford University Press.

Lezak, M. D. (1995) *Neuropsychological Assessment* (3rd ed.). New York: Oxford University Press.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Marcopulos, B. A., McLain, C. A., & Giuliano, A. J. (1997). Cognitive impairment or inadequate norms: A study of healthy, rural, older adults with limited education. *The Clinical Neuropsychologist, 11*(2), 111-131.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). Oxford, England: Oxford University Press.

Mooney, C. Z. (1997). *Monte Carlo simulation.* Thousand Oaks, CA: Sage.

Nelson, H. E., & O'Connell, A. (1978). Dementia: The estimation of premorbid intelligence levels using the New Adult Reading Test. *Cortex, 14*, 234-244.

Nunnally, J. C. (1978). *Psychometric theory.* New York, NY: McGraw-Hill.

Olm-Madden, T. (2008). *A reliable approach to psychological assessment using cognitive test batteries* (Unpublished doctoral dissertation). University of Southern Queensland, Toowoomba, Australia.

Pallant, J. (2010). *SPSS: Survivor manual* (4th ed.). Berkshire, England: Open University Press.

Payne, R. W., & Jones, G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology, 13*, 115-121.

Rael, L. T., Chelune, G. J., Taylor, M. J., Woodward, T. S., Heaton, R. K. (2006). Development of demographic norms for four new WAIS-III/WMS-III indexes. *Psychological Assessment, 18*, 174-181.

Reitan, R.M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation.* Tucson, AZ: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

Rey, A. (1964). *L'examen clinique en psychologie.* Paris: Presses Universitaires de France.

Richardson, E. D., & Marottoli, R. A. (1996). Educational-specific normative data on common neuro-psychological indices for individuals older than 75 years. *Clinical Neuropsychologist, 10*(4), 375-381.

Russell, E. W., Russell, S. L. K., & Hill, B. D. (2005). The fundamental psychometric status of neuropsychological batteries. *Archives of Clinical Neuropsychology, 20*, 785-794.

Sheskin, D. J. (2004). *Handbook of parametric and nonparametric statistical procedures* (3rd ed.). Florida: Chapman & Hall/CRC.

Spreen, O., & Strauss, E. (1991). *A compendium of neuropsychological tests.* NY: Oxford University Press.

Strauss, E., Sherman, E. M. S. & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms and commentary* (3rd ed.).

New York, NY: Oxford University Press.
The Psychological Corporation. (1997). *WAIS-III/WMS-III Technical Manual.* San Antonio, TX: Author.
Wechsler, D. (1981). *Manual of the Wechsler Adult Intelligence Scale – Revised.* San Antonio, TX: The Psychological Corporation.
Wechsler, D. (1997a). *Manual for the Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.
Wechsler, D. (1997b). *Wechsler Memory Scale – 3rd Edition (WMS-III).* San Antonio, TX: The Psychological Corporation.
Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition.* San Antonio, TX: Pearson.
Western Psychological Services. (1983). *The Hooper Visual Organisation Test manual* (Rev. ed.). Los Angeles: Western Psychological Services.
Williams, J. M., & Cottle, C. C. (2011). A correction for recruitment bias in norms derived from meta-analysis. *Psychological Assessment, 23,* 856-864.
Willner, A. E. (1971). *Conceptual Level Analogy Test.* New York: Cognitive Testing Service.

APPENDIX A:

Normalised Standard Scores Calculations

TMT A
<50 years; <12 years education

Raw Score	Frequency	Cumulative Frequency
14.00	2	2.7
15.00	1	4.1
15.18	1	5.5
15.69	2	8.2
16.00	1	9.6
16.70	1	11.0
17.00	2	13.7
17.16	1	15.1
17.37	1	16.4
18.22	1	17.8
18.30	2	20.5
18.66	1	21.9
19.00	1	23.3
19.29	1	24.7
19.30	1	26.0
19.50	1	27.4
19.53	2	30.1
19.80	1	31.5
19.90	1	32.9
20.00	2	35.6
20.81	1	37.0
21.40	1	38.4
21.50	1	39.7
21.60	1	41.1
21.70	1	42.5
21.75	1	43.8
21.90	1	45.2
22.00	2	47.9
22.07	1	49.3
23.00	2	52.1
23.09	1	53.4
23.15	1	54.8
23.20	1	56.2
23.22	1	57.5
23.30	1	58.9
23.44	1	60.3
24.00	2	63.0

TMT B
<50 years; <12 years education

Raw Score	Frequency	Cumulative Frequency
31.94	1	1.4
33.00	1	2.8
33.80	1	4.2
34.85	1	5.6
36.00	1	6.9
36.63	1	8.3
38.00	1	9.7
38.38	1	11.1
38.57	1	12.5
39.16	1	13.9
39.50	1	15.3
40.00	1	16.7
40.09	1	18.1
40.49	1	19.4
42.00	1	20.8
43.80	1	22.2
44.00	1	23.6
44.07	1	25.0
46.00	1	26.4
46.90	1	27.8
48.00	1	29.2
48.47	1	30.6
48.90	1	31.9
49.63	1	33.3
49.75	1	34.7
50.00	1	36.1
50.40	1	37.5
50.47	1	38.9
51.00	1	40.3
51.52	1	41.7
51.55	1	43.1
51.77	1	44.4
52.20	1	45.8
53.09	1	47.2
53.30	1	48.6
55.65	1	50.0
56.00	1	51.4

Raw Score	Frequency	Cumulative Frequency
24.25	1	65.8
24.29	1	67.1
24.56	1	68.5
24.60	1	69.9
24.72	1	71.2
24.83	1	72.6
25.06	1	74.0
25.09	1	75.3
25.96	1	76.7
26.00	1	78.1
27.00	1	79.5
27.62	1	80.8
29.00	1	82.2
29.47	1	83.6
29.56	1	84.9
31.00	2	87.7
33.91	1	89.0
37.00	2	91.8
37.30	1	93.2
37.34	1	94.5
38.50	1	95.9
41.10	1	97.3
42.81	1	98.6
49.12	1	100.0
Total	73	100

Raw Score	Frequency	Cumulative Frequency
56.20	1	52.8
57.00	1	54.2
57.45	1	55.6
57.94	1	56.9
59.00	1	58.3
59.10	1	59.7
59.61	1	61.1
59.87	1	62.5
60.00	2	65.3
64.00	1	66.7
65.00	1	68.1
65.68	1	69.4
66.85	1	70.8
69.97	1	72.2
70.00	2	75.0
70.68	1	76.4
71.00	1	77.8
73.00	2	80.6
75.00	1	81.9
76.96	1	83.3
78.90	1	84.7
79.00	1	86.1
80.00	1	87.5
81.00	1	88.9
86.91	1	90.3
95.00	1	91.7
96.56	1	93.1
110.00	1	94.4
115.52	1	95.8
117.00	1	97.2
120.00	1	98.6
160.00	1	100.0
Total	72	100

TMT A
<50 years; 12+ years education

Raw Score	Frequency	Cumulative Frequency
11.62	1	.4
12.00	1	.8
12.40	1	1.2
12.46	1	1.6
12.50	1	2.0
12.63	1	2.4
12.80	1	2.8
13.00	1	3.2
13.10	1	3.6
13.26	1	4.0
14.00	3	5.2
14.40	1	5.6
14.47	1	6.0
15.00	4	7.6
15.04	1	8.0
15.10	2	8.8
15.20	1	9.2
15.25	1	9.6
15.30	1	10.0
15.33	1	10.4
15.50	1	10.8
15.69	1	11.2
15.80	1	11.6
16.00	3	12.7
16.03	1	13.1
16.59	1	13.5
16.75	1	13.9
17.00	2	14.7
17.20	1	15.1
17.30	1	15.5
17.40	1	15.9
17.50	1	16.3
17.74	1	16.7
18.00	7	19.5
18.22	2	20.3
18.28	1	20.7
18.30	1	21.1

TMT B
<50 years; 12+ years education

Raw Score	Frequency	Cumulative Frequency
15.84	1	.4
22.19	1	.8
23.31	1	1.2
24.50	1	1.6
25.00	1	2.0
26.00	1	2.4
27.00	1	2.8
27.43	1	3.2
28.00	1	3.6
28.40	1	4.0
29.00	1	4.4
29.25	1	4.8
29.70	1	5.2
31.00	3	6.4
31.75	1	6.8
32.30	1	7.2
32.41	1	7.6
32.66	1	8.0
33.00	1	8.4
33.96	1	8.8
33.97	1	9.2
34.00	2	10.0
36.00	2	10.8
36.10	1	11.2
36.24	1	11.6
36.56	1	12.0
37.00	2	12.8
37.11	1	13.2
37.22	1	13.6
37.70	1	14.0
38.50	2	14.8
38.80	1	15.2
39.00	3	16.4
39.25	1	16.8
39.40	1	17.2
39.59	1	17.6
39.80	1	18.0

Raw Score	Frequency	Cumulative Frequency
18.40	1	21.5
18.46	1	21.9
18.50	1	22.3
18.52	1	22.7
18.53	1	23.1
18.69	1	23.5
18.80	2	24.3
18.85	1	24.7
18.91	2	25.5
18.97	1	25.9
19.00	6	28.3
19.07	1	28.7
19.09	1	29.1
19.20	2	29.9
19.24	1	30.3
19.35	1	30.7
19.40	1	31.1
19.50	1	31.5
19.60	1	31.9
19.87	1	32.3
20.00	3	33.5
20.13	1	33.9
20.20	2	34.7
20.22	1	35.1
20.28	1	35.5
20.59	1	35.9
20.66	1	36.3
20.70	1	36.7
20.77	1	37.1
21.00	7	39.8
21.03	1	40.2
21.16	1	40.6
21.25	2	41.4
21.30	1	41.8
21.35	1	42.2
21.50	1	42.6
22.00	7	45.4
22.03	1	45.8
22.13	1	46.2
22.25	1	46.6
22.28	1	47.0
22.30	1	47.4

Raw Score	Frequency	Cumulative Frequency
40.00	3	19.2
40.24	1	19.6
40.38	1	20.0
40.40	1	20.4
40.99	1	20.8
41.00	1	21.2
41.22	1	21.6
41.50	1	22.0
41.70	1	22.4
41.81	1	22.8
42.00	4	24.4
42.28	1	24.8
42.68	1	25.2
42.80	1	25.6
42.90	1	26.0
43.00	2	26.8
43.03	1	27.2
43.08	1	27.6
43.19	1	28.0
43.38	1	28.4
43.51	1	28.8
43.70	1	29.2
43.80	1	29.6
44.00	1	30.0
44.13	1	30.4
44.81	1	30.8
45.00	3	32.0
45.06	1	32.4
45.15	1	32.8
45.38	1	33.2
45.42	1	33.6
45.44	1	34.0
45.62	1	34.4
45.70	1	34.8
45.75	1	35.2
46.00	1	35.6
46.25	1	36.0
47.00	2	36.8
47.25	1	37.2
47.60	1	37.6
47.66	1	38.0
47.72	1	38.4

Raw Score	Frequency	Cumulative Frequency
22.80	1	47.8
22.84	1	48.2
22.90	1	48.6
23.00	4	50.2
23.60	1	50.6
23.63	1	51.0
23.72	1	51.4
23.89	1	51.8
23.90	1	52.2
24.00	7	55.0
24.03	1	55.4
24.06	1	55.8
24.09	1	56.2
24.21	1	56.6
24.53	1	57.0
24.64	1	57.4
24.73	1	57.8
24.82	1	58.2
25.00	8	61.4
25.09	2	62.2
25.25	1	62.5
25.28	1	62.9
25.79	1	63.3
25.83	1	63.7
25.91	1	64.1
26.00	2	64.9
26.10	1	65.3
26.22	1	65.7
26.31	1	66.1
26.40	1	66.5
26.43	1	66.9
26.47	1	67.3
26.69	1	67.7
26.72	1	68.1
27.00	5	70.1
27.16	1	70.5
27.63	1	70.9
28.00	4	72.5
28.10	1	72.9
28.40	1	73.3
28.63	1	73.7
28.66	1	74.1
29.00	3	75.3

Raw Score	Frequency	Cumulative Frequency
48.00	1	38.8
48.21	1	39.2
48.24	1	39.6
48.88	1	40.0
48.99	1	40.4
49.00	7	43.2
49.30	1	43.6
49.48	1	44.0
49.81	1	44.4
49.82	1	44.8
50.00	2	45.6
50.16	1	46.0
50.20	1	46.4
50.57	1	46.8
50.75	2	47.6
50.97	1	48.0
51.00	3	49.2
51.30	1	49.6
51.69	1	50.0
51.70	1	50.4
51.77	1	50.8
52.00	3	52.0
52.22	1	52.4
52.26	1	52.8
52.28	1	53.2
52.90	1	53.6
53.00	3	54.8
53.30	1	55.2
53.33	1	55.6
53.40	1	56.0
53.44	1	56.4
53.72	1	56.8
53.90	1	57.2
54.00	3	58.4
54.53	1	58.8
54.56	1	59.2
55.00	3	60.4
55.20	1	60.8
56.00	2	61.6
56.10	1	62.0
56.44	1	62.4
56.79	1	62.8
56.94	1	63.2

Raw Score	Frequency	Cumulative Frequency
29.13	1	75.7
29.40	1	76.1
29.41	1	76.5
29.56	1	76.9
29.91	1	77.3
30.00	4	78.9
30.04	1	79.3
30.20	1	79.7
30.24	1	80.1
30.28	1	80.5
30.53	1	80.9
30.57	1	81.3
30.60	1	81.7
30.86	1	82.1
30.96	1	82.5
31.00	1	82.9
31.59	1	83.3
31.88	1	83.7
32.00	3	84.9
32.50	1	85.3
32.56	1	85.7
33.00	1	86.1
33.19	1	86.5
33.25	1	86.9
33.39	1	87.3
33.52	1	87.6
34.00	2	88.4
35.00	1	88.8
35.19	1	89.2
35.62	1	89.6
35.96	1	90.0
36.00	2	90.8
36.24	1	91.2
36.85	1	91.6
36.95	1	92.0
37.85	1	92.4
38.00	1	92.8
38.72	1	93.2
38.95	1	93.6
39.00	2	94.4
39.04	1	94.8
40.10	1	95.2

Raw Score	Frequency	Cumulative Frequency
57.00	3	64.4
57.54	1	64.8
58.00	4	66.4
58.06	1	66.8
58.26	1	67.2
58.64	1	67.6
58.70	1	68.0
58.78	1	68.4
59.00	2	69.2
59.09	1	69.6
59.22	1	70.0
59.72	1	70.4
59.75	1	70.8
59.79	1	71.2
59.94	1	71.6
60.00	3	72.8
60.16	1	73.2
60.43	1	73.6
61.00	1	74.0
61.87	1	74.4
62.00	3	75.6
62.60	1	76.0
63.00	1	76.4
63.56	1	76.8
64.00	3	78.0
65.00	4	79.6
65.28	1	80.0
65.56	1	80.4
65.60	1	80.8
65.72	1	81.2
66.00	1	81.6
66.46	1	82.0
66.69	1	82.4
67.00	1	82.8
67.22	1	83.2
67.69	1	83.6
68.00	2	84.4
68.27	1	84.8
69.00	3	86.0
69.27	1	86.4
69.44	1	86.8
70.00	1	87.2

Raw Score	Frequency	Cumulative Frequency
43.00	1	95.6
43.30	1	96.0
43.94	1	96.4
45.00	1	96.8
46.21	1	97.2
46.41	1	97.6
47.00	1	98.0
49.00	1	98.4
49.20	1	98.8
49.43	1	99.2
49.72	1	99.6
56.66	1	100.0
Total	251	100

Raw Score	Frequency	Cumulative Frequency
71.00	1	87.6
71.28	1	88.0
72.06	1	88.4
73.00	1	88.8
74.00	1	89.2
75.00	1	89.6
75.31	1	90.0
76.00	1	90.4
76.30	1	90.8
76.59	1	91.2
77.38	1	91.6
78.00	1	92.0
78.99	1	92.4
79.00	1	92.8
81.97	1	93.2
82.00	1	93.6
83.00	2	94.4
86.00	1	94.8
89.37	1	95.2
94.30	1	95.6
98.74	1	96.0
102.00	1	96.4
106.85	1	96.8
107.00	1	97.2
108.00	1	97.6
111.00	1	98.0
113.00	1	98.4
114.00	1	98.8
115.78	1	99.2
120.00	1	99.6
174.00	1	100.0
Total	251	100

TMT A
50 + years; <12 years education

Raw Score	Frequency	Cumulative Frequency
6.00	1	2.6
18.00	1	5.1
20.00	1	7.7
22.00	2	12.8
23.50	1	15.4
24.00	1	17.9
25.53	1	20.5
26.00	1	23.1
26.60	1	25.6
27.25	1	28.2
28.00	1	30.8
28.01	1	33.3
29.00	2	38.5
29.94	1	41.0
30.00	1	43.6
31.50	1	46.2
31.69	1	48.7
32.00	2	53.8
32.03	1	56.4
32.97	1	59.0
33.00	1	61.5
36.00	3	69.2
38.00	1	71.8
40.00	1	74.4
42.00	1	76.9
42.68	1	79.5
43.00	2	84.6
48.00	1	87.2
49.00	1	89.7
51.00	1	92.3
60.00	1	94.9
61.00	1	97.4
62.00	1	100.0
Total	39	100

TMT B
50 + years; <12 years education

Raw Score	Frequency	Cumulative Frequency
30.00	1	2.6
37.93	1	5.3
43.00	1	7.9
50.00	1	10.5
56.00	1	13.2
58.29	1	15.8
59.34	1	18.4
62.00	3	26.3
62.65	1	28.9
64.53	1	31.6
64.72	1	34.2
67.00	1	36.8
67.75	1	39.5
67.78	1	42.1
70.00	1	44.7
73.00	1	47.4
74.09	1	50.0
75.00	1	52.6
77.00	1	55.3
78.00	3	63.2
86.00	1	65.8
87.00	2	71.1
88.00	1	73.7
89.63	1	76.3
90.69	1	78.9
91.00	1	81.6
91.41	1	84.2
92.00	1	86.8
105.00	1	89.5
127.00	1	92.1
136.00	1	94.7
144.00	1	97.4
183.00	1	100.0
Total	39	100

TMT A
50 + years; 12+ years education

Raw Score	Frequency	Cumulative Frequency
7.00	1	1.9
17.00	1	3.8
18.00	1	5.7
20.00	1	7.5
21.00	1	9.4
22.00	1	11.3
23.00	4	18.9
23.11	1	20.8
23.71	1	22.6
24.00	5	32.1
24.41	1	34.0
25.00	2	37.7
25.10	1	39.6
26.78	1	41.5
27.00	1	43.4
28.00	3	49.1
29.00	1	50.9
29.06	1	52.8
29.22	1	54.7
29.56	1	56.6
30.00	4	64.2
31.00	1	66.0
31.38	1	67.9
31.50	1	69.8
32.00	1	71.7
32.50	1	73.6
33.15	1	75.5
33.18	1	77.4
34.00	2	81.1
36.00	1	83.0
37.00	1	84.9
39.00	1	86.8
40.00	1	88.7
40.88	1	90.6
43.00	1	92.5
45.00	2	96.2
49.00	1	98.1
67.00	1	100.0
Total	53	100

TMT B
50 + years; 12+ years education

Raw Score	Frequency	Cumulative Frequency
29.00	1	1.9
32.00	1	3.8
35.00	1	5.7
37.00	1	7.5
41.00	1	9.4
43.00	1	11.3
49.00	1	13.2
49.94	1	15.1
50.00	1	17.0
51.00	2	20.8
54.09	1	22.6
55.00	3	28.3
56.53	1	30.2
58.00	1	32.1
59.63	1	34.0
60.00	1	35.8
62.00	2	39.6
63.86	1	41.5
64.00	2	45.3
64.50	1	47.2
65.00	2	50.9
67.00	2	54.7
68.00	1	56.6
69.00	1	58.5
71.00	2	62.3
72.91	1	64.2
73.00	1	66.0
74.00	1	67.9
75.00	1	69.8
76.00	2	73.6
79.00	1	75.5
80.00	2	79.2
81.00	1	81.1
83.00	1	83.0
83.50	1	84.9
83.96	1	86.8
88.78	1	88.7
92.00	1	90.6
93.43	1	92.5
110.00	1	94.3

Raw Score	Frequency	Cumulative Frequency
125.00	1	96.2
126.00	1	98.1
130.00	1	100.0
Total	53	100